

FEDERAL UNIVERSITY OF ESPÍRITO SANTO
TECHNOLOGY CENTER
GRADUATE PROGRAM IN ELECTRICAL ENGINEERING

HAMILTON RIVERA FLOR

**DEVELOPMENT OF A MULTISENSORIAL SYSTEM FOR EMOTION
RECOGNITION**

VITÓRIA

2017

HAMILTON RIVERA FLOR

**DEVELOPMENT OF A MULTISENSORIAL SYSTEM FOR EMOTION
RECOGNITION**

Dissertation submitted to the Graduate Program
in Electrical Engineering from the Technology
Center of the Federal University of Espírito
Santo, as partial requirement for obtaining
Master's Degree in Electrical Engineering.

Advisor: Prof. Dr. Teodiano Freire Bastos Filho

VITÓRIA

2017

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)

F632d Flor, Hamilton Rivera, 1983-
Development of a multisensorial system for emotion
recognition / Hamilton Rivera Flor. – 2017.
113 f. : il.

Orientador: Teodiano Freire Bastos.
Dissertação (Mestrado em Engenharia Elétrica) –
Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Expressão facial. 2. Olhos – Movimentos. 3. Kinect
(Controlador programável). 4. Sistemas de computação
interativos. 5. Interação homem-máquina. 6. Variação térmica
facial. I. Bastos, Teodiano Freire. II. Universidade Federal do
Espírito Santo. Centro Tecnológico. III. Título.

CDU: 621.3

HAMILTON RIVERA FLOR

**DEVELOPMENT OF A MULTISENSORIAL SYSTEM FOR EMOTION
RECOGNITION**

Dissertation submitted to the Graduate Program
in Electrical Engineering from the Technology
Center of the Federal University of Espírito
Santo, as partial requirement for obtaining
Master's Degree in Electrical Engineering

Date Defended/Approved: 17 March, 2017. Vitória -ES

Prof. Dr. Teodiano Freire Bastos Filho
Advisor

Profa. Dra. Olga Regina Pereira Bellon
Federal University of Paraná (UFPR)

Profa. Dra. Eliete Maria de Oliveira Caldeira
Electrical Engineering Department, Federal University of Espírito Santo

VITÓRIA

2017

ABSTRACT

Automated reading and analysis of human emotion has the potential to be a powerful tool to develop a wide variety of applications, such as human-computer interaction systems, but, at the same time, this is a very difficult issue because the human communication is very complex. Humans employ multiple sensory systems in emotion recognition. At the same way, an emotionally intelligent machine requires multiples sensors to be able to create an affective interaction with users. Thus, this Master thesis proposes the development of a multisensorial system for automatic emotion recognition.

The multisensorial system is composed of three sensors, which allowed exploring different emotional aspects, as the eye tracking, using the IR-PCR technique, helped conducting studies about visual social attention; the Kinect, in conjunction with the FACS-AU system technique, allowed developing a tool for facial expression recognition; and the thermal camera, using the FT-RoI technique, was employed for detecting facial thermal variation. When performing the multisensorial integration of the system, it was possible to obtain a more complete and varied analysis of the emotional aspects, allowing evaluate focal attention, valence comprehension, valence expressions, facial expression, valence recognition and arousal recognition.

Experiments were performed with sixteen healthy adult volunteers and 105 healthy children volunteers and the results were the developed system, which was able to detect eye gaze, recognize facial expression and estimate the valence and arousal for emotion recognition,

This system also presents the potential to analyzed emotions of people by facial features using contactless sensors in semi-structured environments, such as clinics, laboratories, or classrooms. This system also presents the potential to become an embedded tool in robots to endow these machines with an emotional intelligence for a more natural interaction with humans.

Keywords: emotion recognition, eye tracking, facial expression, facial thermal variation, integration multisensorial.

RESUMO

A leitura e análise automatizada da emoção humana tem potencial para ser uma ferramenta poderosa para desenvolver uma ampla variedade de aplicações, como sistemas de interação homem-computador, mas, ao mesmo tempo, é uma questão muito difícil porque a comunicação humana é muito complexa. Os seres humanos empregam múltiplos sistemas sensoriais no reconhecimento emocional. Assim, esta dissertação de mestrado propõe o desenvolvimento de um sistema multissensorial para reconhecimento automático de emoções.

O sistema multisensorial é composto por três sensores, que permitiram a exploração de diferentes aspectos emocionais, o seguimento do olhar, utilizando a técnica IR-PCR, ajudou a realizar estudos sobre atenção social visual; O Kinect, em conjunto com a técnica do sistema FACS-AU, permitiu o desenvolvimento de uma ferramenta para o reconhecimento da expressão facial; E a câmera térmica, usando a técnica FT-RoI, foi empregada para detectar a variação térmica facial. Ao realizar a integração multissensorial do sistema, foi possível obter uma análise mais completa e variada dos aspectos emocionais, permitindo avaliar a atenção focal, a compreensão da valência, a expressão da valência, a expressão facial, o reconhecimento de valência e o reconhecimento de excitação.

Experimentos foram realizados com dezesseis voluntários adultos saudáveis e 105 crianças saudáveis e os resultados foram o sistema desenvolvido, capaz de detectar o foco do olhar, reconhecer expressões faciais e estimar a valência e a excitação para o reconhecimento emocional.

Este sistema também apresenta o potencial para analisar as emoções das pessoas por características faciais usando sensores sem contato em ambientes semi-estruturados, como clínicas, laboratórios ou salas de aula. Este sistema também apresenta o potencial de se tornar uma ferramenta incorporada em robôs para dotar essas máquinas de uma inteligência emocional para uma interação mais natural com os seres humanos.

List of Figures

| | |
|---|----|
| Figure 1.1 Affective computing advocates the idea of emotionally intelligent machines that can recognize and simulate emotions. | 15 |
| Figure 1.2 The emotions are mental processes present since childhood and they are very important in human communication. | 16 |
| Figure 1.3 Different applications for automatic emotion recognition: A) The integration of emotion recognition with robotic rehabilitation, B) Monitoring drowsiness or attentive state and emotional status of the driver, C) emotion recognition with computer serious games into a rehabilitation scenario, D) robots with social skills. | 17 |
| Figure 1.4 Computer applications and robots used as a pedagogical tool for the social development of children with autism. | 18 |
| Figure 1.5 MARIA is a robot to stimulate cognitive and social interaction skills in children with ASD, A) MARIA (2013-2015), B) New-MARIA (2015-2018) | 19 |
| Figure 1.6 MARIA 2 blocks diagram. | 20 |
| Figure 1.7 Timeline for the evolution of emotion recognition. | 22 |
| Figure 2.1 Experimental platform implemented | 32 |
| Figure 2.2 Software diagram | 36 |
| Figure 2.3 EyeTribe UI Interface | 37 |
| Figure 2.4 Brekel Pro Face interface for kinect 2 | 38 |
| Figure 2.5 Therm-App Android interface. | 39 |
| Figure 2.6 UltraVNC interface for client-server remote control. | 39 |
| Figure 2.7 Environment for experimental test. | 40 |
| Figure 2.8 Real environment for experimental tests. | 40 |
| Figure 3.1 General system of the eye tracker. | 42 |
| Figure 3.2 A) Eye tracking process; B) Diagram block of the IR-PCR eye tracker system. | 44 |
| Figure 3.3 Blocks diagram of the proposed eye tracking interface. | 44 |
| Figure 3.4 Server communication for eye-tracker data acquisition. | 45 |
| Figure 3.5 Experimental tests to optimize operating set-up point. | 47 |
| Figure 3.6 Analysis and graphic report. A- Topographic image of eye tracker data, B- Time eye tracking data graphic and C- spatial eye tracking data graphic. | 48 |
| Figure 3.7 Graphic User Interface. | 48 |
| Figure 3.8 Class diagram of the developed eye tracking interface. | 49 |
| Figure 3.9 Set up for the experimental tests. | 49 |

| | |
|---|----|
| Figure 3.10: Original and filtered signal of the eye tracker, output signal (red) and filtered signal (blue). | 50 |
| Figure 3.11: Histogram for different images..... | 51 |
| Figure 3.12: Example of superposition A) superposition of data and image; B) superposition of topographic eye tracker and image..... | 52 |
| Figure 3.13 GUI Application: A) Robot command using eye tracking, B) Emotion recognition using eye tracking. | 52 |
| Figure 4.1 Expressions recognition system using kinect. | 54 |
| Figure 4.2 Six basic facial expressions describer by Paul Ekman | 55 |
| Figure 4.3. Feature-based Automatic Facial Action Analysis (AFA) system (source: Ying-Li, 2001) | 60 |
| Figure 4.4 Block diagram of the proposed Expression recognition system | 61 |
| Figure 4.5 Kinect data acquisition: A) Depth image; B) Infra-red image; C) Color image. | 61 |
| Figure 4.6 Face detection and 3D facial model creation. | 62 |
| Figure 4.7 Module for AU feature extraction..... | 63 |
| Figure 4.8: Experimental procedure. Participants imitating the model of emotion facial expression displayed on the screen. | 66 |
| 38 Figure 4.9: Emotional facial expressions viewed by the participants. E1 (surprise), E2 (sadness), E3 (anger), E4 (disgust), E5 (fear) and E6 (happiness) (Source: Du 2014). | 66 |
| Figure 4.10 Twenty AUs signals obtained from eight different volunteers imitating the six basic expressions | 67 |
| Figure 5.1 System used to study facial thermal variation detection..... | 70 |
| Figure 5.2 Techniques for thermal variation detection. A) Facial Thermal - Region of Interest (FT-RoI); B) Facial Thermal Feature Points (FTFP). (Source: Salazar-López 2015)..... | 72 |
| Figure 5.3 Example of Facial Thermal – Region of Interest (FT-RoI). | 73 |
| Figure 5.4 FTFPs on human face, facial muscle map, and a geometric profile of the FTFPs. (Source: KHAN, 2006). | 74 |
| Figure 5.5 block diagram of the system here developed for thermal facial variation detection..... | 74 |
| Figure 5.6 Example of the thermal image acquisition process..... | 75 |
| Figure 5.7 Example of the RoIs segmentation. | 76 |
| Figure 5.8 Features used in this work (RoIs temperature, RoI –BaseLine, and RoI - Facial Temperature)..... | 76 |
| Figure 5.9 Features extraction (RoI - BaseLine and RoI - Facial Temperature)..... | 77 |
| Figure 5.10 Thermal images for the six facial expressions considered in this work..... | 78 |
| Figure 5.10 Thermal images for negative, neutral and positive valence..... | 78 |
| Figure 5.11 Thermal images for low, medium and high arousal..... | 79 |
| Figure 6.1 Multisensorial integration: Thermal-Camera-Kinect-Eye Tracker..... | 81 |

| | |
|--|----|
| Figure 6.2 Block diagram of the proposed integration strategy | 83 |
| Figure 6.3 Data-level integration online for data processing. | 84 |
| Figure 6.4 Eye tracker- Kinect integration: A) Focus of attention detection, B) Facial expression recognition..... | 85 |
| Figure 6.5 Calibration process for AUs projection on the thermal image..... | 86 |
| Figure 6.6 Projection of AU points from Color to Thermal image: A) Facial expression detection; B) Facial thermal variation; C) Integration of AUs on thermal image..... | 86 |
| Figure 6.7 Multisensorial integration: A) Focal attention detection, B) Facial expression recognition, C) Estimation of emotional state..... | 87 |
| Figure 7.1 Stimuli used in the three experiments conducted in this research; A) images for valence study; B) names of the basic emotions; C) emotion-inducing videos; D) pictures relative to human facial expressions. | 88 |
| Figure 7.2 A) Setup for the experimental tests. B) Set de images for Valence Study. Source: IAPS (Lang 2008)..... | 89 |
| Figure 7.3 Examples of human face emotional expressions used in the procedure 2. (Source: Du, 2014) | 91 |
| Figure 7.4: Data from eye-tracking sensor referent to attention focus, featured by blue circles overlapping on the facial image. The mean focus obtained is shown in red square. | 92 |

List of Tables

| | |
|--|----|
| Table 1.1: Functional and technical requirements of the project..... | 20 |
| Table 1.2 Modalities for emotions recognition. | 25 |
| Table 1.3 Techniques for emotions recognition. | 28 |
| Table 1.4 Summary of the characteristics of publicly accessible emotional databases. | 30 |
| Table 2.1 Eye tracker EyeTribe features (Source: Theeyetribе, 2013). | 33 |
| Table 2.2 Kinect device features (Source: Kinect for Windows, 2014)..... | 34 |
| Table 2.3 Therm-App features (Source: Therm-App, 2014)..... | 35 |
| Table 2.4 characteristics of emotional data bases implemented in this work..... | 41 |
| Table 3.1: Errors for off-set test..... | 50 |
| Table 3.2: Velocity of tracking test. | 50 |
| Table 3.3: concentric window size test. | 51 |
| Table 3.4: Command rate test. | 51 |
| Table 4.1 Upper face action units and some combinations (source: Ying-Li, 2001). | 57 |
| Table 4.2 Lower face action units and some combinations (source: Ying-Li, 2001). | 57 |
| Table 4.3 Action Units list in FACS system (Source: Ekman 1982). | 58 |
| Table 4.4. Multi-state facial component models of a lip (source: Ying-Li, 2001)..... | 59 |
| Table 4.5. Description of the 20 AU features detected in this system. | 63 |
| Table 4.6 Accuracy of the emotion recognition for three class..... | 67 |
| Table 4.7 Confusion matrix for six emotion recognition using LDA | 68 |
| Table 4.8 Confusion matrix for six emotion recognition using KNN | 68 |
| Table 5.1 Muscular alignment of FTFPs. (Source: KHAN, 2006). | 74 |
| Table 5.2 Percentage of RoIs thermal variation for facial expressions | 78 |
| Table 5.3 Percentage of RoIs thermal variation in arousal and valence..... | 79 |
| Table 7.1: Percentage of the time of viewing of the pictures. | 90 |
| Table 7.2. Number of observers who present highest and lowest attention to pictures featured by the valence. | 90 |
| Table 7.3 Mean and standard deviation of the focus points performed by the participants during the visualization of facial expressions..... | 92 |
| Table 7.4 Time to recognize the emotional facial expressions. | 92 |
| Table 7.5 Number of mistakes in the facial expressions recognition. | 93 |
| Table 7.6 Values for expression recognition. | 93 |

| | |
|--|----|
| Table 7.7 Emotions that each video is intended to evoke. | 94 |
| Table 7.8 Recognition of emotions evoked for each video by volunteer 1.1 | 95 |
| Table 7.9 Recognition of emotions evoked for each video by volunteer 2. | 95 |
| Table 7.10 Recognition of emotions evoked for each video by volunteer 3. | 96 |
| Table 7.11 Validation of functional and technical sensors features | 98 |

List of Abbreviations and Acronyms

| | |
|--------|--|
| ANS | Autonomous Nervous System |
| ASD | Autism Spectrum Disorder |
| AU | Action Unit |
| ECG | ElectroCardioGraphy |
| EEG | ElectroEncephaloGraph |
| EOG | ElectroOculoGraphy |
| EMG | ElectroMyoGraphy |
| FACS | Facial Action Coding System |
| FAP | Face Animation Parameters |
| FE | Facial Expression |
| FER | Facial Expression Recognition |
| FTFP | Facial Thermal Feature Points |
| FT-RoI | Facial Thermal - Region of Interest |
| HRV | Heart Rate Variability |
| IRTI | Infrared Thermal Imaging |
| MAX | Maximally Discriminative Facial Movement |
| MPEG | Moving Pictures Experts Group |
| TC | Thermal Camera |
| TIV | Thermal Intensity Values |

Contents

| | |
|--|----|
| ABSTRACT | 5 |
| 1. INTRODUCTION..... | 15 |
| 1.1. Motivation | 16 |
| 1.2. Context of the problem..... | 19 |
| 1.3. State of the art..... | 21 |
| 1.4. Objectives..... | 31 |
| 1.5. Organization of the document | 31 |
| 2. METHODOLOGY | 32 |
| 2.1. Experimental Platform..... | 33 |
| 2.1.1. Hardware | 33 |
| 2.1.2. Software | 36 |
| 2.2. Environment for experimental test | 40 |
| 2.3. Procedures | 41 |
| 2.4. Database..... | 41 |
| 3. EYE GAZE POINT DETECTION THROUGH THE EYE TRACKER DEVICE..... | 42 |
| 3.1. Background: Human-Computer interaction using eye tracking strategies..... | 43 |
| 3.2. Implementation of the eye tracker interface | 44 |
| 3.2.1. Data acquisition and management module..... | 45 |
| 3.2.2. Operating set up point calibration and control module | 45 |
| 3.2.3. Analysis and graphic reports module | 47 |
| 3.2.4. Graphic User Interface (GUI) | 48 |
| 3.3. Analysis and results for the ET ToolBox developed..... | 49 |
| 3.4. Discussion..... | 53 |
| 4. FACIAL EXPRESSION RECOGNITION USING THE KINECT | 54 |
| 4.1 Background: facial expression recognition using FACS and AU system | 55 |
| 4.1.1 Facial Action Coding System (FACS) | 56 |
| 4.1.2 Automatic facial features extraction and AU recognition..... | 58 |
| 4.1.3 Facial feature extraction | 59 |
| 4.1.4 Facial expression classification | 59 |
| 4.2 Implementation of the system for expression recognition..... | 60 |
| 4.2.1 Data acquisition | 61 |

| | | |
|--------|---|-----|
| 4.2.2 | Face feature extraction: Action Units (AUs) | 62 |
| 4.2.3 | Expression recognition | 64 |
| 4.3 | Analysis and results | 66 |
| 4.4 | Discussion | 68 |
| 5. | EMOTION DETECTION USING THERMAL CAMERA | 70 |
| 5.1. | Background: application of thermography to study of emotions | 71 |
| 5.2. | Implementation | 74 |
| 5.3. | Analysis and results | 77 |
| 5.4. | Discussion | 79 |
| 6. | MULTISENSORIAL INTEGRATION | 81 |
| 6.1. | Background Multisensorial Integration | 82 |
| 6.2. | Implementation of a multisensorial system for emotion recognition. | 83 |
| 6.2.1. | Data-Level Integration on Processing Language | 83 |
| 6.2.2. | Decision-Level Integration eye tracker and Kinect..... | 84 |
| 6.2.3. | Feature-Level integration Kinect and thermal camera..... | 85 |
| 6.2.4. | Hybrid-Level Integration: eye tracker, Kinect and thermal camera..... | 87 |
| 7. | VALIDATION OF MULTISENSORIAL SYSTEM | 88 |
| 7.1. | Experiment 1: Social Focal Attention Recognition | 89 |
| 7.1.1. | Experimental Protocol | 89 |
| 7.1.2. | Results..... | 90 |
| 7.2. | Experiment 2: Expression comprehension and recognition | 91 |
| 7.2.1. | Experimental Protocol | 91 |
| 7.2.2. | Results..... | 92 |
| 7.3. | Experiment 3: Multisensorial Emotion Analysis | 94 |
| 7.3.1. | Experimental Protocol | 94 |
| 7.3.2. | Results..... | 94 |
| 7.4. | Discussion..... | 96 |
| 8. | CONCLUSIONS AND FUTURE WORKS | 99 |
| | REFERENCES | 104 |

CHAPTER 1

1. INTRODUCTION

Humans employ rich emotional communication channels during social interaction by modulating their speech utterances, facial expressions, and body gestures. They also rely on emotional cues to resolve the semantics of received messages. Interestingly, humans also communicate emotional information when interacting with machines. They express affects and respond emotionally during human-machine interaction. However, machines, from the simplest to the most intelligent ones devised by humans, have conventionally been completely oblivious to emotional information. This reality is changing with the advent of affective computing. Affective computing advocates the idea of emotionally intelligent machines. Hence, these machines can recognize and simulate emotions (figure 1.1). In this context, the purpose of this master thesis is the study and implementation of different affective computing techniques to develop a multisensorial system for emotions recognition.

This chapter exposes the motivation of this thesis, the proposed system and a general introduction of the state of art about automatic emotion recognition (historical development, modalities and the principal techniques). The research objectives are also presented.

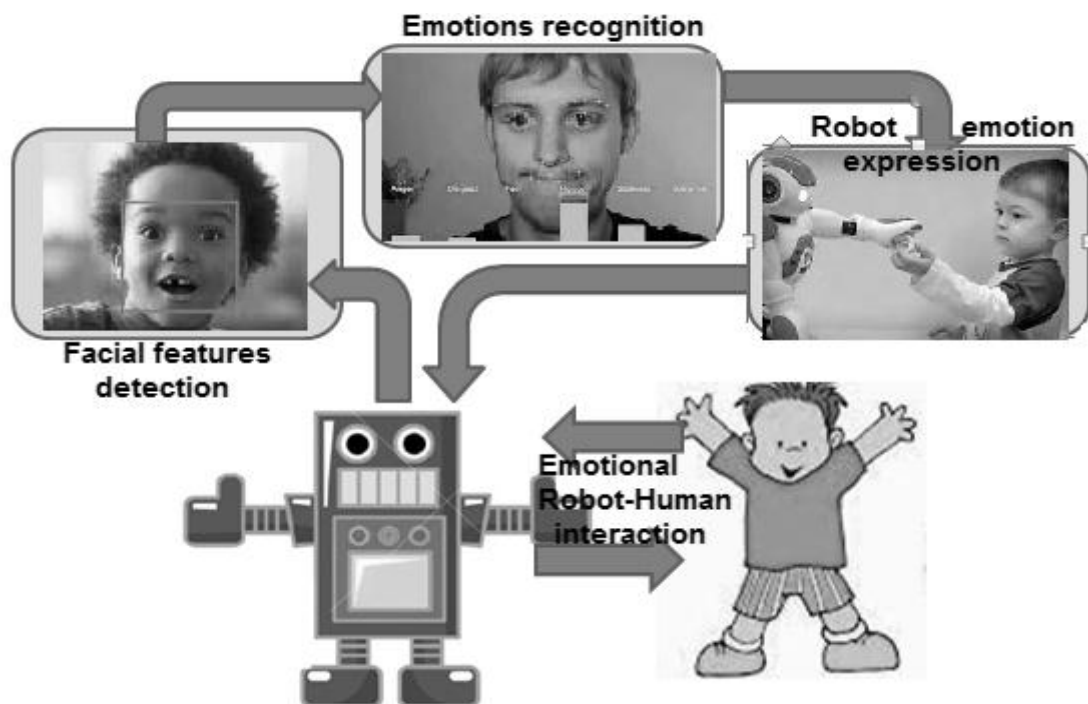


Figure 1.1 Affective computing advocates the idea of emotionally intelligent machines that can recognize and simulate emotions.

1.1. Motivation

To start, it is important to define some basic concepts about emotions. In Hook (2008) emotions are defined as: “strong, rush and relatively unstable mental processes which are followed by some events. Often, emotions are directed to subject that caused it”.

The simplest way to divide emotions is categorize it as negative, positive or neutral (Ekman 1968). In a set of negative emotions are situated, for example, sadness, anger and fear. The second set (positive emotions) contains emotions such as happiness and positive surprise. The last one (neutral category). However, there is another, very important emotion category, namely basic emotions which were defined by Ekman (2003). In his research, he discovered that emotion expression depends only on part from human derivation who identified six basic emotions: anger, sadness, happiness, surprise, fear and disgust.

The ability to recognize and express these emotions has been developing in the process of evolution from thousands of years, therefore such ability is completely natural for human being. Which is present since childhood (Figure 1.2) and allows to take appropriate decisions and have different reactions. Moreover, ability to express emotions allows to notify surrounding people about our mental state. An easy-to-understand example of the processes that generate emotions based on the research of Ekman (2003) is present in the animated movie “Inside Out”, which is about how five emotions (personified as the characters Anger, Disgust, Fear, Sadness and Joy) grapple for control of the mind of an 11-year-old girl named Riley during the tumult of a move from Minnesota to San Francisco.



Figure 1.2 The emotions are mental processes present since childhood and they are very important in human communication.

The motivation for many researchers to study the automatic analysis of human affective behavior is the potential wide variety of applications (Figure 1.3) such as human-computer interaction, health-care, computer assisted learning, anomalous event detection, and interactive computer games. Among various cues that express human emotion, nonverbal information like facial cues, plays an important role in analyzing human behavior. Human emotion recognition, usually combined with speech, gaze and standard interactions, like mouse movements and keystrokes, can be used to build adaptive environments by detecting the user's affective states (Maat and Pantic, 2007). Similarly, one can build socially aware systems (DeVault et al., 2014), or robots with social skills like Sony's AIBO and ATR's (Robovie Ishiguro, 2001). Detecting students' frustration can help improve e-learning experiences (Kapoor, 2007). Gaming experience can also be improved by adapting difficulty, music, characters or mission according to the player's emotional responses (Blom, 2014). Pain detection is used for monitoring patient progress in clinical settings (Irani, 2015). Detection of truthfulness or potential deception can be used during police interrogations or job interviews (Ryan, 2009). Monitoring drowsiness or attentive state and emotional status of the driver is critical for the safety and comfort of driving (Vural 2007). Depression recognition from facial expressions is a very important application in the analysis of psychological distress (Scherer 2013). Finally, in recent years successful commercial applications like Emotient, Affectiva, RealEyes and Kairos perform largescale internet-based assessments of viewer reactions to ads and related material for predicting buying behavior.



Figure 1.3 Different applications for automatic emotion recognition: A) The integration of emotion recognition with robotic rehabilitation, B) Monitoring drowsiness or attentive state and emotional status of the driver, C) emotion recognition with computer serious games into a rehabilitation scenario, D) robots with social skills.

Robotic for rehabilitation and therapy has established a new paradigm for higher efficiency and physical performance compared to the frequently tedious conventional rehabilitation process based on the repetition principle stated in Burke et al. (2009). The integration of a robot with computer serious games into a rehabilitation or therapy scenario has outlined a promising approach by offering sessions in a more stimulating physical and psychological re-education environment. Furthermore, rehabilitation robotic provides a repository for data analysis, diagnosis, therapy customization and maintenance of patient records. The involvement of the user is probably one of the most important mechanisms through which therapy produces clinical benefits. At the same time the engagement of the user with therapeutic exercises is an important topic in the rehabilitation robotics research field.

Sophisticated software and robots are currently being implemented in pedagogical therapies aiming at the behavioral improvement of children with Autism Spectrum Disorder (ASD). (Figure 1.4). Applications for intervening in emotional and social recognition skills are presented by Thomeer et al. (2015) and Scassellati et al. (2012). In the literature, there are examples of robots with playful friendly form, used as a pedagogical tool for the social development of children with autism (Michaud e Clavet, 2001; Robins et al., 2005; Goulart et al., 2015). These robots are meant to get the child's attention and to stimulate him/her to interact with the environment. In addition, they provide situations of significant and sophisticated interaction through speech, sounds, visual indications and movements (Michaud e Clavet, 2001). Facilitating contact and visual focus, the robot can be a platform for shared interaction, allowing other people (other children with or without autism and adults) to interact instantly. Thus, robots can facilitate the interaction of children with ASD with other humans (Robins et al., 2005; Werry et al., 2001).



Figure 1.4 Computer applications and robots used as a pedagogical tool for the social development of children with autism.

Studies aimed at improving and understanding the behavior and emotions of individuals with ASD are increasing due to the improvement in the technological area, the development of increasingly robust computers and robots, and better sensors. Thus main motivation in the development of this research is to contribute with the understanding of emotions in the therapy of children with and without autism.

1.2. Context of the problem

Between the years 2013 and 2015 at the Intelligent Automation Laboratory of the Federal University of Espírito Santo (UFES-LAI) the robot MARIA (acronym for Mobile Autonomous Robot for Interaction with Autistics) was built. MARIA is a mobile robot with a special costume and a monitor to display multimedia contents, designed to stimulate cognitive and social interaction skills in children with ASD, promoting eye gaze, touch, and imitation, besides interaction with people. Figure 1.5A shows an image of the first version of the robot MARIA, which was developed by Goulard (2015) and Valadão (2016). Although the usability of this robot was demonstrated, it has some limitations, such as the fact of the robot be only remotely controlled, and not having a emotion recognition system onboard. This pilot studies with MARIA showed the need to create a new version of this robot named New-MARIA (Figure 1.5B), in order to include new devices to catch the children's attention, it enhance the probability of interaction with children with ASD, both in terms of quantity and quality.

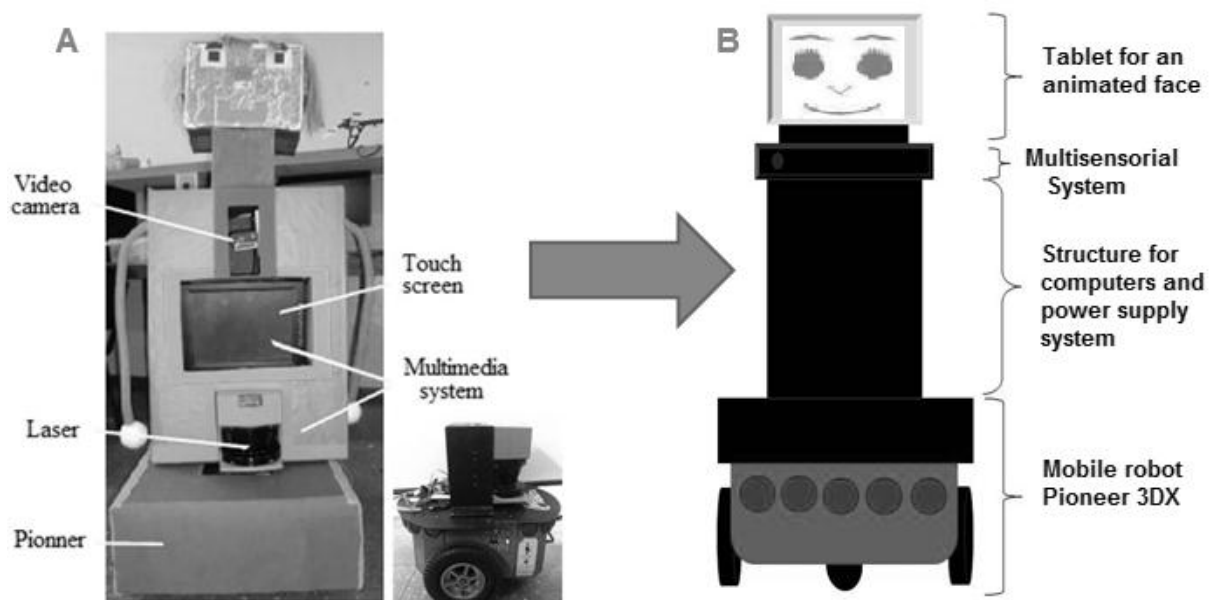


Figure 1.5 MARIA is a robot to stimulate cognitive and social interaction skills in children with ASD, A) MARIA (2013-2015), B) New-MARIA (2015-2018)

New-MARIA, is still in development, but uses a system of cameras and sensors capable of capturing images of children with ASD, to identify classes of emotions and focus on an object or an image. It also has an animated face for interaction with such children. Those new features were designed in order to facilitate the stimulation of social skills and study of emotions and focus of attention.

In the New-MARIA project, five sub-systems were proposed, which allow autonomous navigation, robot control, multimedia interaction, social interaction, therapeutic-Robot-Child approach and automatic emotion recognition. Figure 1.6 shows the block diagram of the sub-system of New-MARIA.

1.3. State of the art

1.3.1 Historical evolution of emotion recognition

The scientific study of the emotions began with Charles Darwin's *The Expression of Emotions in Man and Animals* book, first published in London in 1872 (Darwin 1872). Darwin gathered evidence that some emotions have an universal facial expression, cited examples and published pictures suggesting that emotions are evident in other animals, and proposed principles explaining why particular expressions occur for particular emotions which, he maintained, applied to the expressions of all animals. Most of such systems attempted to recognize a small set of prototypic emotional expressions, i.e. joy, surprise, anger, sadness, fear, and disgust. Following the work of Darwin (1872) more recently Ekman (1976, 1993) and Izard et al. (1983) proposed that basic emotions have corresponding prototypic facial expressions.

Recognizing user's emotional state is then one of the main requirements for computers to successfully interact with humans. Most of the works in the affective computing field do not combine different modalities into a single system for the analysis of human emotional behavior, different channels of information (mainly facial expressions and speech) are considered independently to each other. In the area of unimodal emotion recognition, there have been many studies using different, but single, modalities. Facial expressions in Pantic (2000), vocal features in Scherer (1996), body movements in Camurri (2003) and McNeill (1992). Unimodal sensors have been used as inputs during these attempts, while multimodal emotion recognition is currently gaining ground (Pantic, 2003). Nevertheless, most of the works consider the integration of information from facial expressions and speech and there are only a few attempts to combine information from body movement and gestures in a multimodal framework. Gunes and Piccardi (2006), for example, fused, at different levels, facial expressions and body gestures information for bimodal emotion recognition. In this study we explore the state of the art about the various modalities used in emotion classification and the most important techniques used for emotions recognition.

First scientific study of the emotion published

- Darwin (1872): study about emotions in "The Expression of the Emotions in Man and Animals"

Emotions have a universal facial expression

- Ekman and Friesen (1976-1993): study about facial emotions and FACS in "Pictures of Facial Affect" and "Facial expression and emotion"
- Lang (1980-1990): study about SAM in "Emotion, attention, and the startle reflex"

Studies about modalities for emotions recognition

- McNeill (1992): study about body gestures in "Hand and mind: What gestures reveal about thought"
- Scherer (1996): study about Speech, in "Adding the Affective Dimension: A new look in speech analysis and synthesis"

- Rimm-Kaufman and Kagan (1996): study about thermal emotions, in “The psychological significance of changes in skin temperature”
- Genno et al. (1997): study about thermal emotions, in “Using facial skin temperature to objectively evaluate sensations”

Automatic unimodal systems for emotions recognition

- Pantic and Rothkrantz (2000): automatic facial emotions recognition, in “Automatic analysis of facial expressions”
- Schuller and Rigoll (2002): automatic speech, in “Recognising interest in conversational speech-comparing bag of frames and supra-segmental features”
- Camurri et al (2003): automatic Body Gestures, in “Recognizing Emotion from Movement: Comparison of Spectator Recognition and Automated Techniques”

Multimodal systems for emotions recognition

- Pantic and Rothkrantz (2003): multimodal emotions, in “Towards an Affect-sensitive Multimodal Human-Computer Interaction”
- Nakasone et al. (2005): EMG, skin conductance, in “Emotion recognition from electromyography and skin conductance”
- Gunes et al. (2006): automated Multimodal emotions, in “Emotion recognition from expressive face and body gestures”

Quality adaptative multimodal and robust systems

- Gupta et al. (2016): quality adaptative multimodal and robust, in “A quality adaptive multimodal affect recognition system for user-centric multimedia indexing”

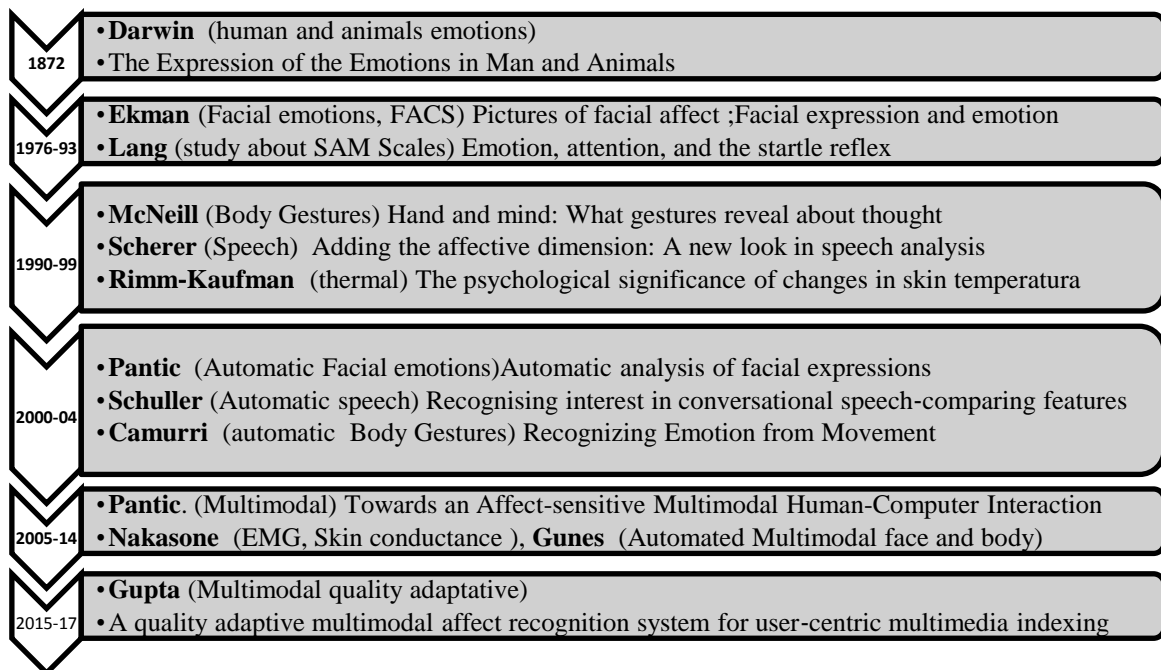


Figure 1.7 Timeline for the evolution of emotion recognition.

1.3.2 Modalities of emotion recognition

Various modalities of emotional channels can be used for the automatic recognition of human emotions and each one provides different measurable information that the machine needs to retrieve and interpret to estimate human emotion.

Visual modality

The visual modality is rich in relevant informational content and includes the facial expression, eye gaze, pupil diameter, blinking behavior, and body expression. The most studied nonverbal emotion recognition method is facial expression analysis (Gelder, 2009). Perhaps, that is because facial expressions are the most intuitive indicators of affect.

An automatic facial analysis system from images or video usually consists of four main parts. First, face detection in the image or face tracking across video frames. Second, feature extraction by many methods requiring a face registration to be performed. During registration, fiducial points (e.g., the corners of the mouth or the eyes) are detected, allowing features extraction from the face with techniques depending on the data modality (Nair, 2009). The approaches are divided into geometric or appearance based, global or local, and static or dynamic. Other approaches use a combination of these categories. Finally, machine learning techniques are used to discriminate between facial expressions. These techniques can predict a categorical expression or represent the expression in a continuous output space, and can either model or do not model temporal information about the dynamics of facial expressions (Alyuz, 2012).

In addition to facial-expression analysis, eye-based features such as pupil diameter, gaze distance, and gaze coordinates, and blinking behavior have been used in multimodal systems. In fact, Panning et al. (2012) found that in their multimodal system, the speech paralinguistic features and eye-blinking frequency were the most contributing modalities to the classification process.

On the other hand, body expression for emotion recognition has been debated in the literature, McNeill (1992) maintains that two-handed gestures are closely associated with the spoken verbs. Hence, they arguably do not present new affective information; they simply accompany the speech modality. Consequently, some researchers Pantic et al. (2003) argue that gestures may play a secondary role in the human recognition of emotions. This suggests that they might be less reliable than other modalities in delivering affective cues that can be automatically analyzed.

Affect recognition using body expression involves tracking the motion of body features in space. Works rely on the use of three-dimensional (3D) measurement systems that require markers to be attached to the subject's body (Kleinsmith et al., 2007). However, some markerless solutions involve video cameras, such as Sanghvi et al. (2011), and wearable sensors, as in Kleinsmith et al. (2011). Once the motion is captured, a variety of features are extracted from body movement. In particular, the following features have been reliably used:

body or body part velocity (Gong et al. 2010), body or body part acceleration (Bernhardt et al., 2007), amount of movement (Savva et al., 2012), joint positions, nature of movement (e.g., contraction, expansion, and upward movement), body parts orientation (e.g., head and shoulder) (Kleinsmith et al., 2007; Savva et al., 2012), and angle or distance between body parts (e.g., distance from hand to shoulder and angle between shoulder to shoulder vectors) (Bernhardt et al., 2007).

Audio modality

Speech carries two interrelated informational channels: linguistic information that express the semantics of the message and implicit paralinguistic information conveyed through prosody. Both of these channels carry affective information.

Linguistic speech channel build an understanding of the spoken message and provides a straightforward way of assessing affect. Typically, an automatic speech recognition algorithm is used to convert speech into a textual message. Then, a sentiment analysis method interprets the polarity or emotional content of the message. However, this approach for affect recognition has its pitfalls because it is not universal, and a natural language speech processor has to be developed for each dialect (Ambady and R. Rosenthal, 1992).

Paralinguistic speech-prosody channel sometimes, it is not about what we say, but how we say it. Therefore, speech-prosody analyzers ignore the meaning of messages and focus on acoustic cues that reflect emotions. Before the extraction of tonal features from speech, preprocessing is often necessary to enhance, denoise, and dereverberate the source signal (Weninger et al., 2015). Then, using windowing functions, low level descriptors (LLDs) features can be extracted, such as: pitch (fundamental frequency F0), energy (e.g., maximum, minimum, and root mean square), linear prediction cepstral (LPC) coefficients, perceptual linear prediction coefficients, cepstral coefficients (e.g., mel-frequency cepstral coefficients, MFCCs), formants (e.g., amplitude, position, and width), and spectrum (mel-frequency and FFT bands) (Eyben et al., 2009).

Physiological modality

Physiological signals can be used for emotion recognition through the detection of biological patterns that are reflective of emotional expressions. These signals are collected through contact sensors that are affixed to the body (Dalglish et al., 2009), and using brain imaging like in Poh et al. (2011) and Mc Duff et al. (2014).

There are a multitude of physiological signals that can be analyzed for affect detection. Typical physiological signals used for the assessment of emotions are electroencephalograph (EEG), electrocardiography (ECG), electromyography (EMG), skin conductance, respiration rate, and skin temperature. In Al Osman and El Saddik (2016) ECG records the electrical activity of the heart and from the ECG signal, the heart rate (HR) and heart rate variability (HRV) can be extracted. HRV is used in numerous studies that assess mental stress (Al Osman, 2016; Healey and Picard, 2005; and Jovanov et al., 2003). EMG measures muscle

activity and is known to reflect negatively valenced emotions (Nakasone, 2005). EEG is the electrical activity of the brain measured through electrodes connected to the scalp and forehead. EEG features are often used to classify emotional dimensions of arousal, valence, and dominance as proposed in (Gupta and Falk, 2016). Skin conductance measures the resistance of the skin by passing a negligible current through the body. The resulting signal is reflective of arousal (Nakasone, 2005) as it corresponds to the activity of the sweat glands and the autonomous nervous system (ANS). Finally, respiration rate tends to reflect arousal (Homma and Masaoka, 2008), while skin temperature carries valence cues (Rimm-Kaufman and Kagan, 1996).

Table 1.2 shows a summary of the modalities for emotions recognition and their main characteristics. The modalities with better performance for the technical of the system requirements are facial expression recognition, eye gaze tracking and thermal variation detection.

Table 1.2 Modalities for emotions recognition.

| Modalities | Channels | measured Feature | 7 Contactless | 8 Portable | 9 Robust operation | 10 easy to set up, calibration |
|---------------------------------------|-----------------|--|----------------------|-------------------|---------------------------|---------------------------------------|
| Facial expression | Visual | FACS, AU, FAP | Yes | Yes | Yes | Yes |
| Eye gaze | Visual | Eye gaze | Yes | Yes | Yes | Yes |
| Body expression | Visual | Body gestures | Yes/Not | difficult | Difficult | Difficult |
| Linguistic speech channel | Audio | Speech recognition | Yes | Yes | Difficult | Difficult |
| Paralinguistic speech-prosody channel | Audio | Speech-prosody recognition | Yes | Yes | Yes | Yes |
| Physiological signals (EEG, EMG, ECG) | Physiological | Electrocardiography Electromyography Electroencephalograph Respiration rate | Not | Difficult | Not | Difficult |
| Skin conductance | Physiological | Skin resistance | Not | Yes | Yes | Difficult |
| Thermal variation | Physiological | Skin temperature | Yes | Yes | Yes | Yes |

After selecting the three modalities that best adapt the requirements (Facial expression. Thermal variation), the techniques to implement each modality are presented below.

1.3.3 Techniques for emotion recognition

Facial expression recognition

Techniques to facial expression recognition may be categorized into two main classes. Descriptive coding schemes parametrize Facial Expressions (FE) in terms of surface properties which focus on what the face can do. Judgmental coding schemes describe FEs in terms of the latent emotions or affectivity that are believed to generate them.

Descriptive coding schemes focus on what the face can do. The most well-known example of such systems are Facial Action Coding System (FACS) that describes all possible perceivable facial muscle movements in terms of predefined action units (AUs). All AUs are numerically coded and facial expressions correspond to one or more AUs. Although FACS is primarily employed to detect emotions, it can be used to describe facial muscle activation regardless of the underlying cause. Face Animation Parameters (FAP), it is a standard to enable the animation of face models defined by the Moving Pictures Experts Group (MPEG) in the MPEG-4 that describes facial feature points (FPs) that are controlled by FAPs. The value of the FAP corresponds to the magnitude of deformation of the facial model in comparison to the neutral state. Though the standard was not originally intended for automated emotion detection, it has been employed for that goal in Lin et al. (2012). These coding systems inspired researchers to develop automated image or video-processing methods that track the movement of facial features to resolve the affective state (Cohen et al., 2003). FAP is now part of the MPEG-4 standard and is used for synthesizing FE for animating virtual faces. It is rarely used to parametrize FEs for recognition purposes (Aleksic and Katsaggelos, 2006). Its coding scheme is based on the position of key feature control points in a mesh model of the face. Maximally Discriminative Facial Movement Coding System (MAX) (Izard, 1983), another descriptive system, is less granular and less comprehensive. Brow raise in MAX, for instance, corresponds to two separate actions in FACS. It is a truly sign-based approach as it makes no inferences about underlying emotions.

Judgmental coding schemes, on the other hand, describe FEs in terms of the latent emotions or affects that are believed to generate them. Because a single emotion or affect may result in multiple expressions, there is no 1:1 correspondence between what the face does and its emotion label. A hybrid approach is used to define emotion labels in terms of specific signs rather than latent emotions or affects. Examples are EMFACS (Emotion FACS) developed by Ekman and Friesen (1983), which scores facial actions relevant for particular emotion displays, and AFFEX system which is used for identifying affect expressions by holistic judgment (Izard 1983). In each system, expressions related to each emotion are defined descriptively. As an example, enjoyment may be defined by an expression displaying an oblique lip-corner pull co-occurring with cheek raise.

Eye track movements

While a large number of different techniques to track eye movements have been investigated in the past, three eye tracking techniques have emerged as the predominant ones and are widely used in research and commercial applications today. These techniques are: (1) videoculography (VOG), video-based tracking using head-mounted or remote visible light video cameras; (2) video-based infrared (IR) pupil-corneal reflection (PCR); and (3) Electrooculography (EOG). While particularly the first two video-based techniques have a lot of properties in common, all techniques have application areas where they are more useful.

VOG presented in Hansen and Majaranta (2012) and Goldberg and Wichansky (2003) is a video-based eye tracking system that relies on off-the-shelf components and video cameras and it can, therefore, be used for developing “eye aware” or attentive user interfaces that do not strictly require accurate gaze point tracking (Hansen and Pece, 2005). In contrast, due to the additional information gained from the IR-pupil corneal reflection, IR-PCR presented in Duchowski (2003); and Bengoechea et al. (2012), provides highly accurate gaze point measurements of up to 0.5° of visual angle and such technique has, therefore, emerged as the preferred one for scientific domains, such as usability studies or gaze-based interaction, and commercial applications, such as in marketing research. Finally, EOG presented by Hori et al. 2006 and Borghetti et al. (2007) has been used for decades for ophthalmological studies as it allows measuring relative movements of the eyes with high temporal accuracy.

Thermal variation detection

Facial Thermal Feature Points (FTFP) is used to detect transitions of emotional states by synthesizing the facial expressions in Sugimoto (2000). The facial thermal changes caused by muscular movement were analyzed. The system used a neutral expression face and a test face. The two faces were geometrically reformed and were compared in order to develop a thermal differential model. Results of this work suggest that it is possible to detect facial temperature changes caused by both the transition of emotional states and physiological changes.

The results further suggested that detected facial temperature changes could help in understanding the transition of emotional states. A combination of visual images, thermal features, and audio signals were employed for recognizing affective states (Yoshitomi et al., 2000). The study examined possibilities of classifying neutral, happy, sad, angry, and surprised faces through integration of visual, thermal, and audio signals. In another attempt, using bio-physiological signals for achieving Automated Facial Expression Classification (AFEC) and Automated Affect Interpretation (AAI) functionality Khan et al. (2004); Khan et al. (2005) used transient facial thermal features from 21 participant faces for AFEC and AAI. Thermal images with normal and pretended expressions of happiness, sadness, disgust, and fear were captured. Facial points that undergo significant thermal changes with a change in expression termed Facial Thermal Feature Points (FTFPs) were identified.

Facial Thermal - Region of Interest (FT-RoI), researches applying this technique to psychological processes, showed that an activity that involves a mental effort can lower the facial temperature: for our body mental activity resembles the stress response, which produces a process of vasoconstriction in the nose. The interesting thing about this study is that the decrease in temperature is not due to a physiological factor, such as, for example, physical activity, but psychological, a stressful task, showing a specific thermographic pattern. This is the key to applying this technique to other psychological processes that contain similar, comparable responses of the nervous system, such as emotions.

When it comes to studying complex emotions, concepts such as arousal (amount of activation that produces a stimulus) or valence (the positive or negative sense of emotion) are basic concepts (Lang, 1995; and Lang, 2005). the idea has been to use thermography as a somatic marker of emotional response, working with the hypothesis that facial thermograms can be used as reliable indicators of emotional parameters. To this end, three different studies are described by Salazar-López et al. (2015). In all of them the participants visualized several sets of images of different types on a computer, while the thermal camera detected the temperature of their face. For data processing, the face was divided into regions of interest (RoI), such as the forehead, tip of the nose, cheeks or orbital area, and the RoI before, during and after the presentation of stimuli.

Table 1.3 shows a comparison between the techniques studied and from which it was concluded that the techniques IR-PCR, FACS-AU, FT-RoI are those that meet the functional and technical requirements to develop the system for emotion recognition. Commercially are many devices, this work proposed the EyeTribe device to implement the IR-PCR technique, Kinect 2.0 for developed the FACS-AU and Therm-App camera for the FT-RoI technique implementation.

Table 1.3 Techniques for emotions recognition.

| Modalities | Techniques | Methods | Functional requirements achieved | Technical requirements achieved |
|--------------------------|------------|---|----------------------------------|---------------------------------|
| Eye tracking | (VOG) | video based tracking using head-mounted or remote color video | 1,2,3 | 7,10 |
| | (IR-PCR) | video-based infrared pupil-corneal reflection | 1,2,3 | 7,8,9,10 |
| | EOG | Electrooculography | 1 | 8 |
| Facial expression | FAP | Face Animation Paramters | 4 | 7,8,10 |
| | MAX | Maximally Discriminative Facial Movement Coding System | 4,5 | 7,8,9,10 |
| | FACS-AU | Facial Action Coding System-Action Units | 4,5 | 7,8,9,10 |
| | EMFACS | Emotion FACS | 4,5,6 | 7,8,9 |
| | AFFEX | affect expressions by holistic judgment | 4,5 | 7,8,9 |
| Facial thermal variation | FTFP | Facial Thermal Feature Points | 4,5 | 7,8,9 |
| | FT-RoI | Facial Thermal - Region of Interest | 5 | 7,8,9,10 |

1.3.4 Multimodal data base

One of the challenges in developing multimodal emotion recognition methods is the need to collect multisensorial data from many different subjects. Also, it is difficult to compare the obtained results with other studies given that the experimental setup is different. Therefore, it is essential to use data bases to produce repeatable and easy-to-compare results, but currently very few multimodal affect databases are publicly available.

The data bases used in emotion recognition are classified into three types: posed, induced, and natural-emotional. For the posed data bases, the subjects are asked to act out a specific emotion while the result is captured. Typically, facial, body expression and speech information are captured in such databases. However, posed databases has some limitations, as they cannot incorporate biosignals; it cannot be guaranteed that posed emotions trigger the same physiological response as spontaneous ones, according to Jerritta et al. (2011). For induced databases, the subjects are exposed to a stimulus (watching a video or picture, listening an audio or receiving a physical stimulus) in a controlled setting, such as laboratory. The stimulus is designed to evoke certain emotions. In some cases, following the stimulus, the subjects are explicitly asked to act out an emotional expression. The eNTERFACE'05 presented by Martin et al. (2006) is an example of such data base. For the natural data bases, the subjects are exposed to a real-life stimulus such as interaction with human or machine and data collection mostly occurs in a noncontrolled environment. Table 1.4 shows some details of emotional data bases.

Table 1.4 Summary of the characteristics of publicly accessible emotional databases.

| Reference | # de subjects | DB type | Modalities | Description | Labeling |
|---------------------------|-----------------------|---------------------|--|--|---|
| UT-Dallas O'Toole 2005 | 284 | Induced | Visual | 5 emotions, 10 minutes emotion inducing videos | Feel trace |
| AAI Roisman (2004) | 60 | Natural | Visual and audio | 6 emotions, Interviewed and asked to describe the childhood experience | Observers judgment |
| VAM (2008) | 19 | Natural | Visual and audio | Valence, activation , Dominance, dimensional labeling | SAM (valence, arousal) Observers judgment |
| NIST Equinox(2005) | 600 | Posed | Thermal Infrared | 3 emotions (smile, frowning and surprise) | N/A |
| IRIS (2006) | 30 | Posed | Thermal | 3 emotions (surprise, laughing and anger) | N/A |
| GEMEP, Bänziger (2012) | 10 | Posed | Visual and audio | 17 emotions | N/A |
| AFEW, Dhall (2012) [109] | N/A(1426 video clips) | Natural | Visual and audio | Six basic emotions + neutral | Expressive keywords from movie subtitles + observers' verification |
| HUMAINE (2007) [115] | Multiple databases | Induced and natural | Visual, audio, and physiological (ECG, skin conductance and temperature, and respiration) | Varies across databases | Observers' judgment + selfassessment |
| MAHNOB-HCI (2012) | 27 | Induced | Visual (face + eye gaze), audio, and physiological (EEG, ECG, skin conductance and temperature, and respiration) | Dimensional and categorical labeling | Selfassessment (SAM for arousal and valence) |
| PhySyQX (2015) [120] | 21 | Natural | Audio and physiological (EEG and near-infrared spectroscopy, NIRS) | Dimensional labeling | SAM (valence, arousal, dominance) plus nine other quality metrics (e.g., naturalness, acceptance) |

1.4. Objectives

GENERAL

The general goal of this work is to implement an emotion recognition system based on facial features using several sensors, in order to improve the accuracy of the system.

SPECIFICS

- Develop a platform that allows to acquire eye movements and facial thermal and color images.
- Implement methods to detect eye gaze points through an eye tracker device.
- Implement methods for facial expression recognition using color camera.
- Develop algorithms for processing images from the thermal camera for emotion detection.
- Implement methods for emotion analysis based on multisensory integration.
- Evaluate the proposed system using statistical index.

1.5. Organization of the document

This master thesis is divided in eight chapters: Chapter 1 exposes the motivation of this work and the state of the art in emotion recognition application, challenges, opportunities and trends of automatic emotion analysis. The objectives of research are also presented. Chapter 2 contains the methodology and materials including the description of the developed platform and the environment for experimental test. Chapter 3 exposes the overview of eye tracking for visual social attention and the methods implemented to detect the eye gaze point through the eye tracker device. Chapter 4 contains the overview about facial analysis for emotion recognition and the methods developed for facial expression recognition using Kinect. Chapter 5 exposes the overview of thermal analysis for emotion variation and the algorithms implemented for processing images from the thermal camera for emotion detection. Chapter 6 provides the overview of the multisensorial emotion recognition and the integration techniques implemented in this work. Chapter 7 provides the experimental protocol to test the multimodal system and evaluate the proposed system using statistical index. In addition, the results and discussion are also presented. Finally, Chapter 8 presents the conclusions and future works.

CHAPTER 2

2. METHODOLOGY

The methodology proposed in this work is an experimental study of the emotions. To develop this study, the following was required: construct a multisensorial platform for data and image collection, implement methods to process the information collected through sensors, design protocols and procedures for experimental testing, create a database with the acquired information (data, image, signals), define target population of test and specify the characteristics of the experimental environment. Figure 2.1 shows a diagram of the experimental platform implemented.

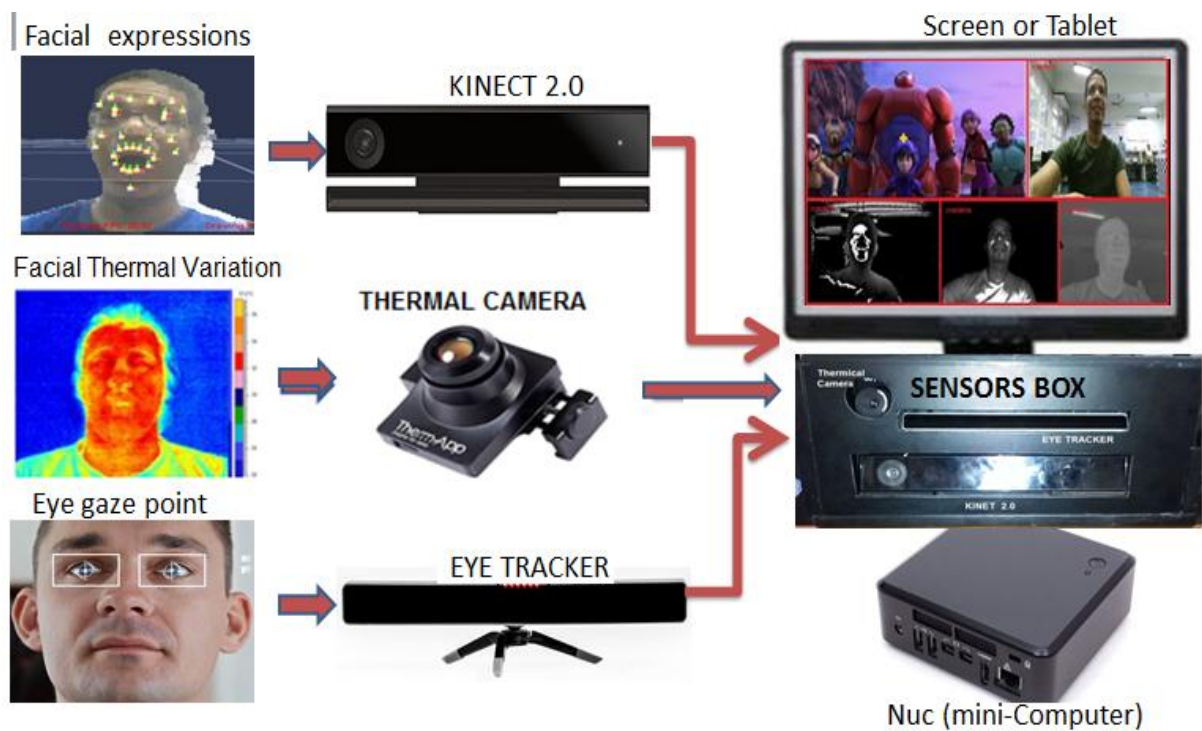


Figure 2.1 Experimental platform implemented

2.1. Experimental Platform

The experimental platform implemented in this work is composed of one Microsoft Kinect 2.0 device, which provides high quality color and depth images to be used to obtain facial points FACS-AU; one Opgal App-therm thermal camera, to be used to determine the RoI (Region of Interest) in the thermal image, so that the features of the face can be extracted; one eye tracker device for gaze tracking; two minicomputer for image and data processing; one tablet for thermal image processing and a Desktop computer, for offline processing and reporting of results. For platform operation different software programs are used.

2.1.1. Hardware

The main equipments used in this work are: the eye tracker (Eye Tribe), the 3D and color camera (Kinect), and the thermal camera (Therm-App).

The Eye Tribe Tracker is an eye tracking system that can calculate the location where a person is looking by means of information extracted from person's face and eyes. The eye gaze coordinates are calculated with respect to a screen the person is looking at, and are represented by a pair of (x, y) coordinates given on the screen coordinate system. The user needs to be located within the tracker's trackbox. The trackbox is defined as the volume in space where the user can theoretically be tracked by the system. When the system is calibrated, the eye tracking software calculates the user's eye gaze coordinates with an average accuracy of around 0.5 to 1° of visual angle. Assuming the user sits approximately 60 cm away from the screen/tracker, this accuracy corresponds to an on-screen average error of 0.5 to 1 cm. The main components of the Eye Tribe Tracker are a camera and a high-resolution infrared LED. Table 2.1 shows the EyeTribe features (Theeyetribe, 2013).

Table 2.1 Eye tracker EyeTribe features (Source: Theeyetribe, 2013).

| | |
|--------------------|--|
| Sampling rate | 30 Hz and 60 Hz mode |
| Accuracy | 0.5° (average) |
| Spatial resolution | 0.1° (RMS) |
| Latency | < 20 ms at 60 Hz |
| Calibration | 5, 9, 12 points |
| Operating range | 45 cm – 75 cm |
| Tracking area | 40 cm × 30 cm at 65 cm distance |
| Screen sizes | Up to 24 inches |
| API/SDK | C++, C# and Java included |
| Data output | Binocular gaze data |
| Dimensions | (W/H/D) 20 × 1.9 × 1.9 cm (7.9 × 0.75 × 0.75 inches) |
| Weight | 70 g |
| Connection | USB 3.0 Superspeed |

Kinect is a line of motion sensing input devices by Microsoft for Xbox 360 and Xbox One video game consoles and Windows PCs. Based around a webcam-style add-on peripheral, it enables users to control and interact with their console/computer without the need for a game controller, through a natural user interface using gestures and spoken commands. Table 2.2 shows the Kinect features (Kinect for Windows, 2014).

Table 2.2 Kinect device features (Source: Kinect for Windows, 2014).

| Feature | Benefits |
|--|--|
| Improved body tracking | The enhanced fidelity of the depth camera, combined with improvements in the software, have led to a number of body tracking developments. The latest sensor tracks as many as six complete skeletons, and 25 joints per person. The tracked positions are more anatomically correct and stable and the range of tracking is broader. |
| Depth sensing 512 x 424 30 Hz FOV: 70 x 60 One mode: .5–4.5 meters | With higher depth fidelity and a significantly improved noise floor, the sensor gives improved 3D visualization, improved ability to see smaller objects and all objects more clearly, and improves the stability of body tracking. |
| 1080p color camera 30 Hz (15 Hz in low light) | The color camera captures full, 1080p video that can be displayed in the same resolution as the viewing screen, allowing for a broad range of powerful scenarios. In addition to improving video communications and video analytics applications, this provides a stable input on which to build high quality, interactive applications. |
| New active infrared (IR) capabilities 512 x 424 30 Hz | In addition to allowing the sensor to see in the dark, the new IR capabilities produce a lighting-independent view—and use IR and color at the same time. |
| Kinect for Xbox One sensor dimensions (length x width x height) | 9.8" x 2.6" x 2.63" (+/- 1/8") 24.9 cm x 6.6 cm x 6.7 cm Length: The Kinect cable is approximately 9.5 feet (2.9 m) long Weight: approximately 3.1 lbs (1.4 kg) |
| A multi-array microphone | Four microphones to capture sound, record audio, as well as find the location of the sound source and the direction of the audio wave. |

Therm-App is an innovative thermal imaging device which offers two image processing modes: superb high-resolution day/night imaging, and basic thermography. Small enough, and combined with a set of interchangeable lenses, Therm-App provides top quality thermal capabilities and the advantages of an open-source platform.

Therm-App extends human vision by turning an Android device into a thermal camera. This lightweight, modular, high resolution device attaches onto Android devices allowing to

display, record, and share thermal images for Night Vision and Thermography applications. Table 2.3 shows the Thermal camera features.

Table 2.3 Therm-App features (Source: Therm-App, 2014).

| | |
|-------------------------------|---|
| Minimal Requirements | Android 4.1 and above, supporting USB OTG |
| Imager | 384 x 288 microbolometer LWIR 7.5 -14um |
| Optics | 6.8mm lens (55° x 41°) 13mm lens (29°x 22°) 19mm lens (19°x14°) 35mm lens (11° x 8°) |
| Focus | Manual, 0.2m to infinity |
| Frame Rate | 8.7Hz |
| Weight | 138 grams / 4.86 ounces |
| Size | 55 x 65 x 40mm (2.16 x 2.55 x 1.57in) |
| Operating Temperature | -10°C to +50°C (14°F to +122°F) |
| Storage Temperature | -20°C to +50°C (-4°F to +122°F) |
| Power Supply | No battery, 5V over USB OTG cable, power consumption < 0.5W |
| Certifications | CE, FCC, RoHS |
| Encapsulation | IP54 |
| Mount/Handle | Ergonomic handle, using 1/4"-20 standard tripod mount |
| Device Attachment | Clip-on for smartphone (5 -10cm span) |
| Resolution | 384 x 288 pixels (>110,000 pixels) |
| Accuracy | +/- 3°C or 3% (@25°C) |
| Sensitivity | NETD <0.07°C |
| Temperature Range Calibration | 5 – 90 °C |

2.1.2. Software

The software implementation of the system requires two operating systems and four programming languages, in addition to various software applications. Figure 2.2 shows a block diagram.

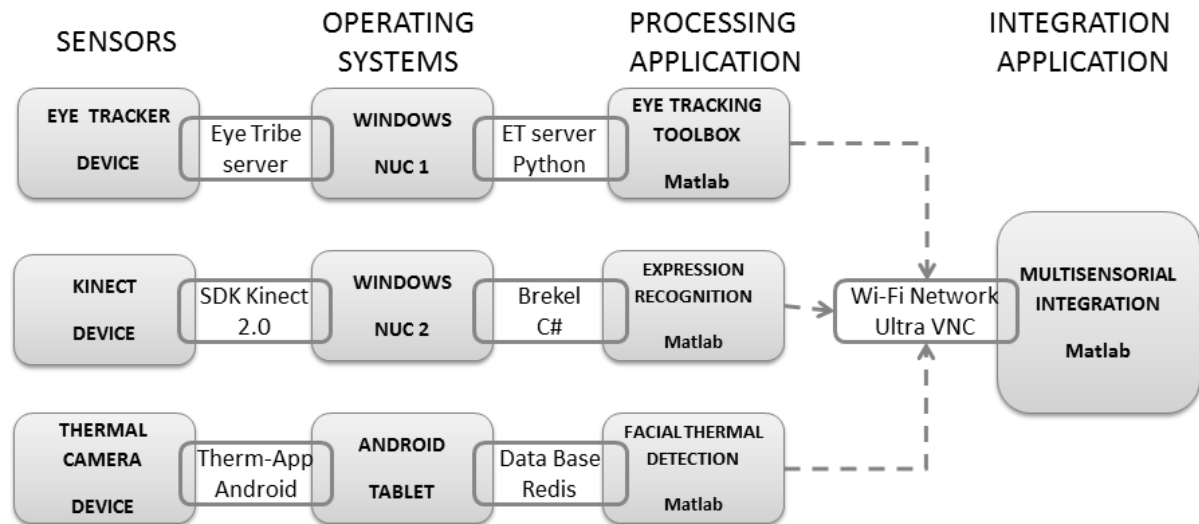


Figure 2.2 Software diagram

The operating system Windows 8.1 and Android Kitkat 4.4 are used. Windows 8.1 is a computer operating system released by Microsoft. Android is a mobile operating system developed by Google, based on the Linux kernel and designed primarily for touchscreen mobile devices such as smartphones and tablets.

The programming languages used in this work are: Processing 3.2.4, Matlab 2013b, C# 6.0 and Python 3.6.0. Processing is a flexible software sketchbook and a language for learning how to code within the context of the visual arts.

Matlab, (Matrix Laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language. A proprietary programming language developed by MathWorks, Matlab allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C#, Java, Fortran and Python.

C# is a multi-paradigm programming language encompassing strong typing, imperative, declarative, functional, generic, object-oriented (class-based), and component-oriented programming disciplines. It was developed by Microsoft within its .NET initiative and later approved as a standard by Ecma (ECMA-334) and ISO (ISO/IEC 23270:2006). C# is one of the programming languages designed for the Common Language Infrastructure. C# is a general-purpose, object-oriented programming language.

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's

elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation. The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language.

The software for eye tracker is the EyeTribe Software Development Kit (SDK) and Python EyeTribe Server version 0.0.3. The EyeTribe SDK is composed of EyeTribe Server and EyeTribe UI. The EyeTribe UI provides a direct feedback of the current tracking state and allows to change the default settings. The main window is depicted in Figure 2.3.

Python EyeTribe Server is an EyeTribe Toolbox for Matlab. It consist on a set of functions that can be used to communicate with eye trackers manufactured by the EyeTribe. The communication process is not direct, but goes via a sub-server that receives input from Matlab (when the functions from this toolbox are called), and then sends commands to the actual EyeTribe server. Its setup is rather odd, but it is the solution to come up with to get around the problem of Matlab not having suitable multithreading functionality. This functionality is required for running a heartbeat Thread (which keeps the connection with the EyeTribe alive), and another Thread to monitor samples (and write these to a log file). Similar results might be obtained by using callback functions within Matlab's TCP/IP framework, but that approach causes timing errors.

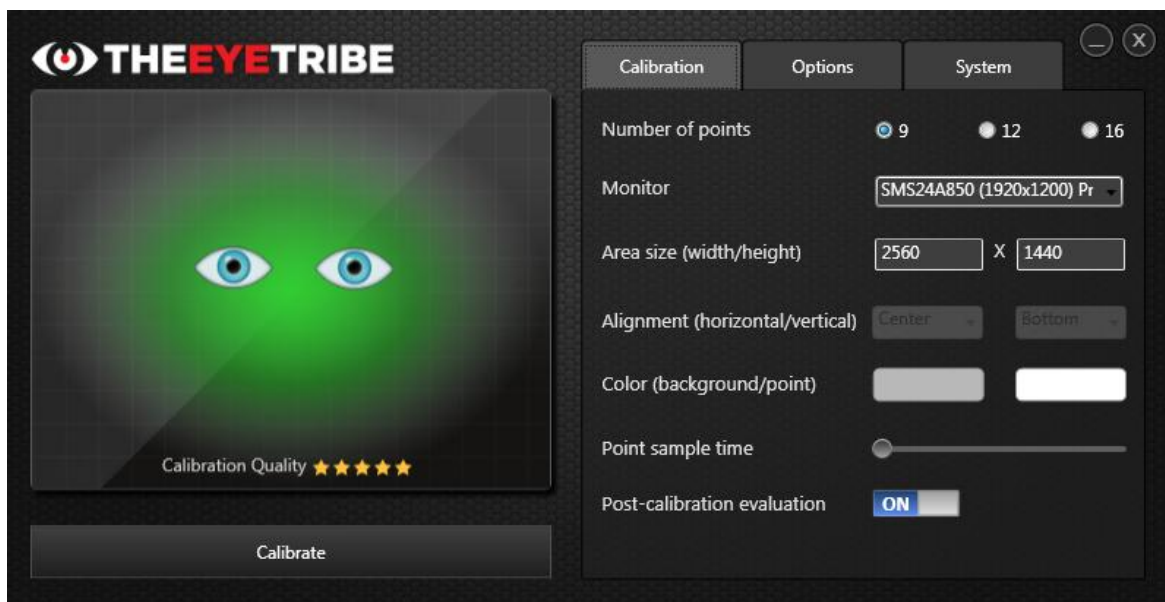


Figure 2.3 EyeTribe UI Interface

Kinect for Windows SDK 2.0.1 allows building desktop applications for Windows. Kinect SDK 2.0 improved body, hand and joint orientation. With the ability to track as many as six people and 25 skeletal joints per person including new joints for hand tips, thumbs, and shoulder center and improved understanding of the soft connective tissue and body positioning, it is able to get more anatomically correct positions for crisp interactions, more accurate avateering, and avatars that are more lifelike. Advanced facial tracking and resolution 20 times greater, enabling the application to create a mesh of more than 1,000 points for a more accurate representation of a person's face. Build avatars that appear more lifelike.

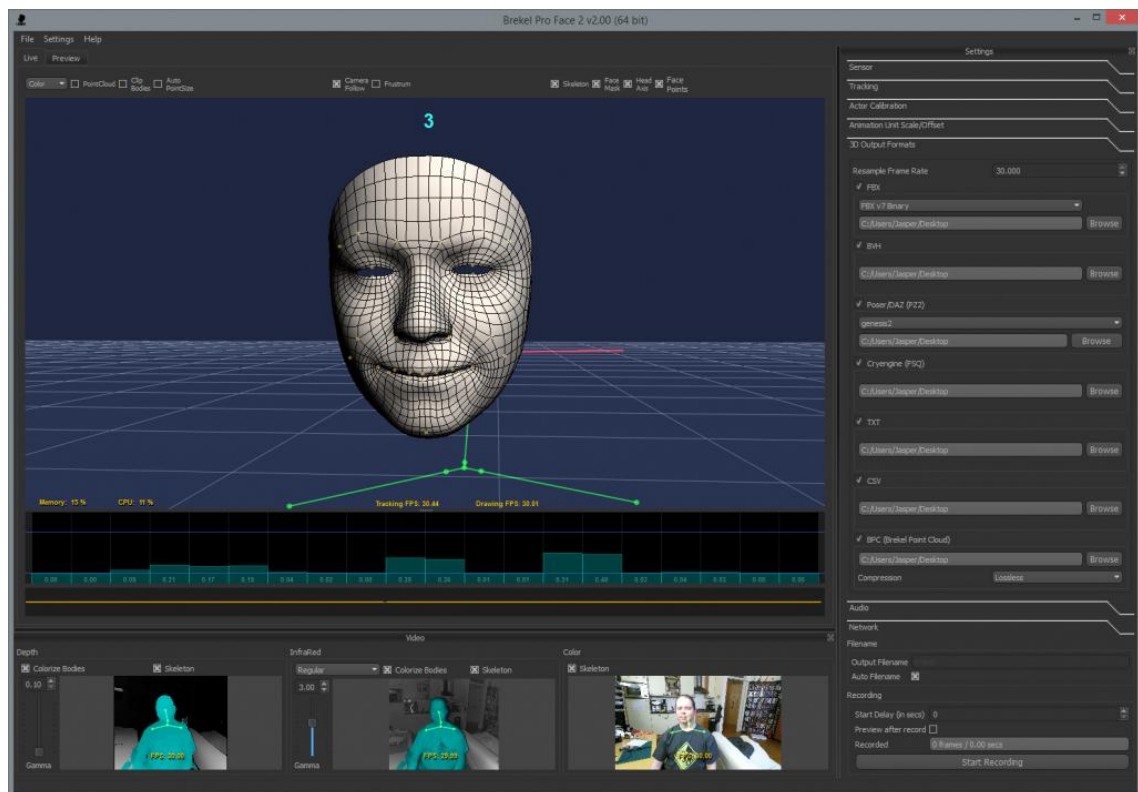


Figure 2.4 Brekel Pro Face interface for Kinect 2

Brekel Pro Face v2 is a Windows application that enables 3D animators to record and stream 3D face tracking of up to 6 people using a Kinect.

Its principal characteristics are multi-person face tracking (1-6 people simultaneously), tracks head position/rotation, tracks 20 different face shapes (including left/right asymmetry), works in realtime, no offline processing required, no calibration required, supports FBX formats v6, v7, ASCII and binary, record audio in sync from Kinect's microphone or any other audio source, adjustable scale/offset per animation unit, build face mesh resembling actor, visualizes Color, InfraRed, Depth, 3D PointCloud and Face Mesh, ability to resample output data from 30fps to custom frame rates and optionally stream directly to the Unity3D game engine.

For the thermal camera are used the software Android Therm-App and Redis BSD 3.2.7. Android Therm-App is an application to use the therm-app camera in Android devices. Redis

is an open source used as a database. It supports data structures such as strings, hashes, lists, sets, sorted sets with range queries, bitmaps, hyperloglogs and geospatial indexes with radius queries. Redis has built-in replication, Lua scripting, LRU eviction, transactions and different levels of on-disk persistence, and provides high availability via Redis Sentinel and automatic partitioning with Redis Cluster.

Temperature Detection Settings

- [1] **Menu**
- [2] **Spot** – Temperature reading on a single spot.
- [3] **Hi/Lo** – Shows the highest and lowest temperatures on the screen.
- [4] **Line** – User draws a line on the screen. The highest and lowest temperatures in the area will be shown along with the average.
- [5] **Area** – User indicates size of a rectangle on the screen. The highest and lowest temperatures in the area are shown along with the average.
- [6] **Temperature alert** – Configure temperature alerts.

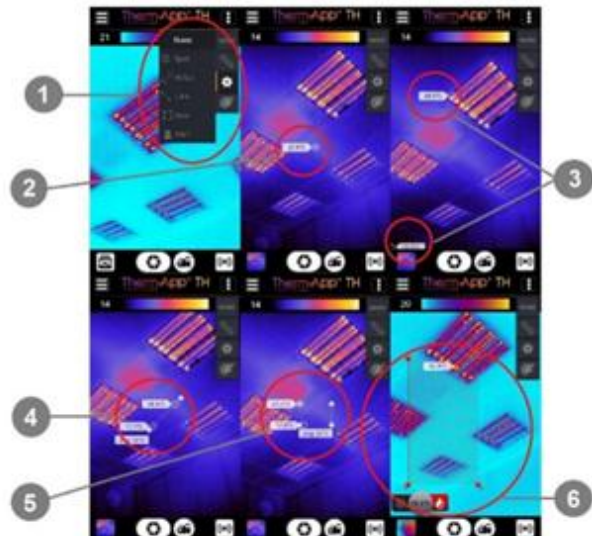


Figure 2.5 Therm-App Android interface.

For communication and remote access a Wi-Fi network and a virtual remote access VNC application are used. Ultra VNC 1.2.1.2 is an easy to use and free remote computer access softwares, that can display the screen of another computer on other screen. VNC use the Remote Frame Buffer protocol (RFB) that allows a desktop to be viewed and controlled remotely over the Internet. A VNC server must be run on the computer sharing the desktop, and a VNC client must be run on the computer that will access the shared desktop. UltraVNC Server and Viewer are an easy to use, free software that can display the screen of one computer (Server) on the screen of another (Viewer). The program allows the viewer to use their mouse and keyboard to control the Server Computer remotely. Figure 2.6 shows the UltraVNC interface.

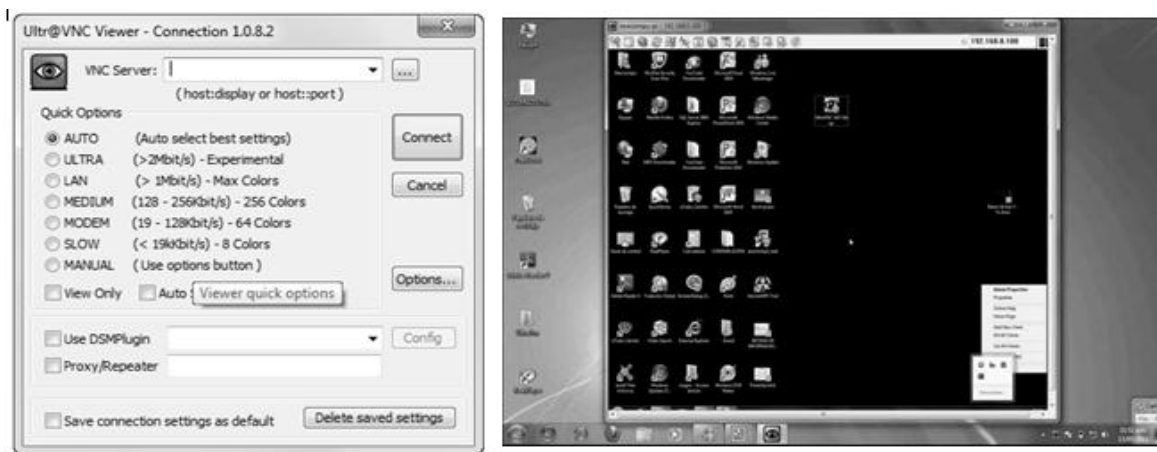


Figure 2.6 UltraVNC interface for client-server remote control

2.2. Environment for experimental test

To set up for the experimental tests, the platform shown in Figure 2.7 is used. Since image and videos are used to induce the subjects' emotion, it is chosen a quiet room as the experimental environment to ensure that the effect of the screened videos is not compromised. The facial emotions recording system includes a color camera system (Kinect), thermal camera, eye tracker, illumination system, thermometer and humidity sensor.

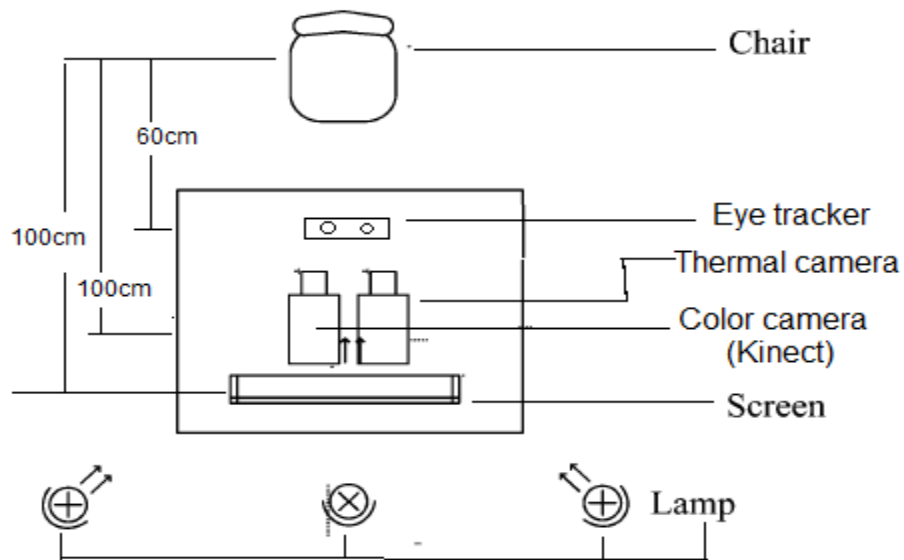


Figure 2.7 Environment for experimental test.

Although thermal emissivity from the facial surface is relatively stable under illumination variations, it is sensitive to the temperature of the environment. Therefore, it is recorded the temperature of the room during the experiments. Room temperature between 20° and 23°, and the humidity ranges between 30% and 40%.



Figure 2.8 Real environment for experimental tests.

2.3. Procedures

This research approaches three procedures in order to assess the visual attention, facial expressions detection and emotion recognition. This work has the approval of the Ethics Committee of UFES, number 1.121.638.

Procedure 1 – this procedure has the participation of sixteen healthy adult volunteers, with mean age of 28 years old (± 5.32). This procedure evaluate social visual attention (focal attention and point of interest of volunteer and valence comprehension).

Procedure 2 – This procedure has the participation of eleven healthy adult volunteers, with mean age of 28.27 years old (± 5.33). This procedure evaluate expression recognition (recognition of facial expressions, focal attention and emotional variation).

Procedure 3 – This procedure has the participation of 105 healthy children volunteers, with age ranged between 6 to 11 years old. This procedure evaluate multisensorial emotion recognition (Focal attention and point of interest of volunteer, facial expression recognition, valence recognition and emotional state).

The experimental procedures use the platform presented in Section 2.1, and environment for experimental test presented in Section 2.2. In Chapter 7, a complete description of the procedures are presented.

2.4. Database

This research proposes a database focused on aspects of posed (MARIA Database 1) and induced (MARIA Database 2) emotion recognition and inference. First, we describe in details the design, collection, and annotation of the database. The number of subjects is 16 adults and 105 children, the modalities of recognition are visual (face + eye gaze) and physiological (thermal). The description of emotion target are six basic emotions, valence and activation, emotion positive or negative. The data labeling is, Observers' verification, SAM (valence, arousal) and Observers judgment. Table 2.4 shows details of emotional database implemented in this work.

Table 2.4 characteristics of emotional data bases implemented in this work.

| Reference | # de subjects | DB type | Modalities | Description |
|------------------|---------------------------|---------|-----------------------------------|--|
| Maria Database 1 | 16 adults 105 Children | Posed | Visual (face + eye gaze), Thermal | 6 basic emotions |
| Maria Database 2 | 16 adults 105 Children | Induced | Visual (face + eye gaze), Thermal | 6 basic emotions, valence and activation |

CHAPTER 3

3. EYE GAZE POINT DETECTION THROUGH THE EYE TRACKER DEVICE.

Eye tracking is a technology that consists of calculating the eye gaze point of a user as he/she looks around. A device equipped with an eye tracker (ET) enables users to use their eye gaze as an input modality that can be combined with other input devices like mouse, keyboard, touch and gesture, referred as active applications in games, operative system navigation, e-books, market research studies, and usability testing. These eye tracker applications can be used for new assistive technologies in medical and psychological research, and in this research there is an interest of studying the use of eye tracker for visual social attention applications.

This chapter exposes the development of a toolbox for Matlab with four modules which allow the use of eye tracking systems for therapy (physical, psychological and medical), control of intelligent environment and studies about visual social attention. The interface developed allows the volunteers to connect applications from Matlab to the eye tracker device, thus allowing them to control the sampling time, to set-up and configure the system, besides that, to manage the eye tracker data. Furthermore, they can generate analysis and graphic reports, and control the graphic interface. This chapter also presents the different kinds of analyses during several types of eye tracker tests (off-set error, velocity of tracking, latency, concentric windows, graphics report and graphic user interface).

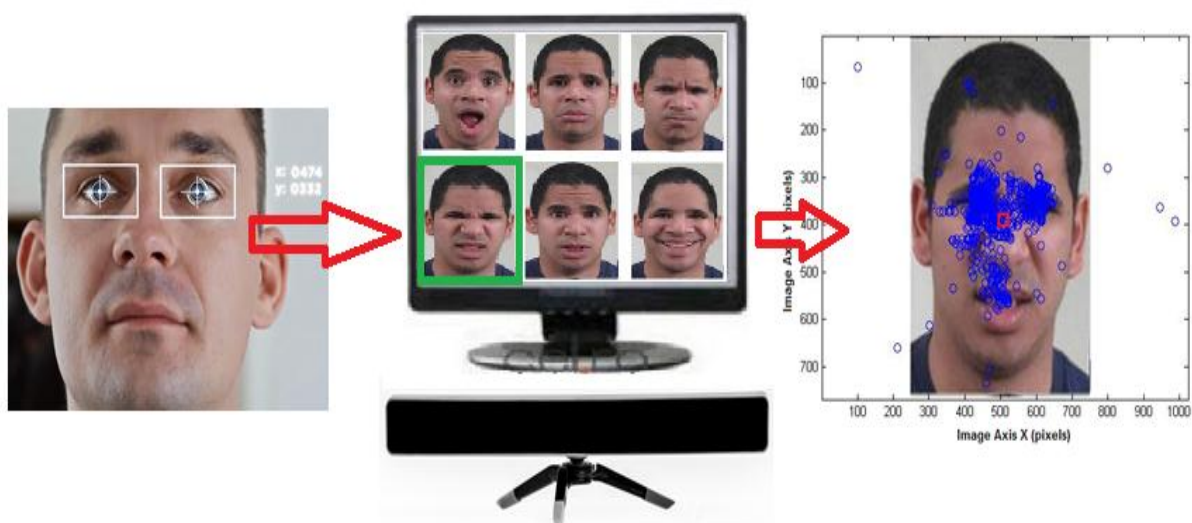


Figure 3.1 General system of the eye tracker.

3.1. Background: Human-Computer interaction using eye tracking strategies

The most publicized strategy of applying eye tracking to existing interfaces is to use the eye to perform pointing and selection tasks. Glenstrup and Engell-Nielsen (1995) have argued on numerous evidences that there is a relationship between the interests of individuals and what they are looking. Another source of evidence is the model built by Card Moran and Newell (1980), which provides the time T spent for the execution of pointing tasks with the use of traditional devices such as the mouse. Equation (3.1) shows a simplified version of the model:

$$T = TM + TV + TP + TR \quad (3.1)$$

In such simplified model, the motor subtask (i.e., what effectively moves the device into the correct position), takes time TP , preceded by a cognitive task that takes TM time, and a visual task (i.e., visual target search in question) that takes time TV . TR is the system response time. Today, despite evidence that this model is not accurate (Hornof and Kieras, 1999), it still provides a reasonable estimate and justify the efforts to apply tracing to look at the execution of pointing tasks.

However, the direct mapping look (more specifically of fixations) to a system selection command creates the problem identified by Jacob (1990), called "Midas Touch", in which a selection can be activated at any position of the observed screen by the user, whether he/she intended to do it or not. This makes necessary a post-filtering for the acquired eye tracker data, representing a challenge in designing interaction technique to avoid the Midas Touch problem and, implementing mechanisms to make the computer understands when the user wants to perform a selection command. The first approach to solve this problem is the implementation of a lag time, in which the selection of an object is performed only after a time interval.

Several applications using eye tracking can be found in literature. Kocejko, Bujnowski, and Wtorek (2008) presented an Eye Mouse, which is a system to people with motor disabilities where the mouse position is controlled by eye gaze. Lupu et al. (2011) presented a system called Asistsys, which is based on eye tracking, making it possible to the people with motor disability to express their wishes and needs only by visualizing options on a monitor.

Studies about the impact of a system based on eye tracking in the quality of life of people with amyotrophic lateral sclerosis, a neurodegenerative disease, are presented in (Calvo et al., 2008). In fact, the eye tracking technique has quite potential of application in interfaces to people with this disease, because they maintain their cognitive ability and, in the most cases, the ability to control the eye gaze. According to Calvo (2008), people who took part in its studies and tested the system noted an improvement in the quality of life, because they were able to communicate independently, and the communication was easier, briefer and less painful.

Figure 3.2 shows a block diagram of the InfraRed Pupil-Corneal Reflection (IR-PCR), eye tracking system, that is considered in this work. The Eye Tribe Tracker is an eye tracking device that can calculate the location where a person is looking by means of information

extracted from person's eye and head. The eye gaze coordinates are calculated with respect to a screen where the person is looking at, and are represented by a pair of (x, y) coordinates given on the screen coordinate system. In order to track the user's eye movements and calculate the on-screen gaze coordinates, the eye tracker must be placed below the screen and pointing to the user.

The user needs to be located within the tracker's trackbox. The trackbox is defined as the volume in space where the user can theoretically be tracked by the system. The size of the trackbox depends on the frame rate, with a higher frame rate offering a smaller trackbox.

When the system is calibrated, the eye tracking software calculates the user's eye gaze coordinates with an average accuracy of around 0.5 to 1° of the visual angle. Assuming the user sits approximately 60 cm away from the screen/tracker, this accuracy corresponds to an on-screen average error of 0.5 to 1 cm.

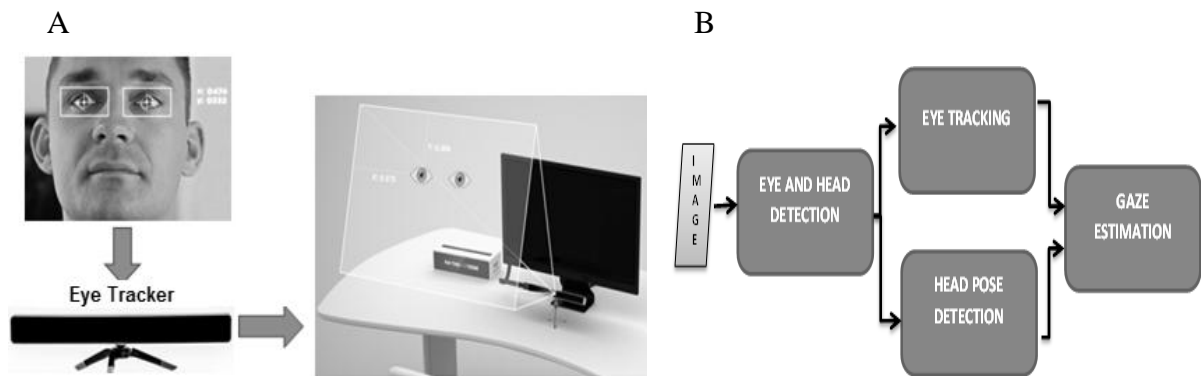


Figure 3.2 A) Eye tracking process; B) Diagram block of the IR-PCR eye tracker system.

3.2. Implementation of the eye tracker interface

Figure 3.3 shows the block diagram of the interface developed in Matlab to facilitate the use of different eye tracker systems in assistive technologies. In this interface four modules are implemented, which allow: acquiring and managing data from the eye tracker (ET); calibrating and preparing the system according to the user disability; analyzing and generating graphical reports; and finally, facilitating the implementation of graphical interfaces controlled by eye gaze using the eye tracker.

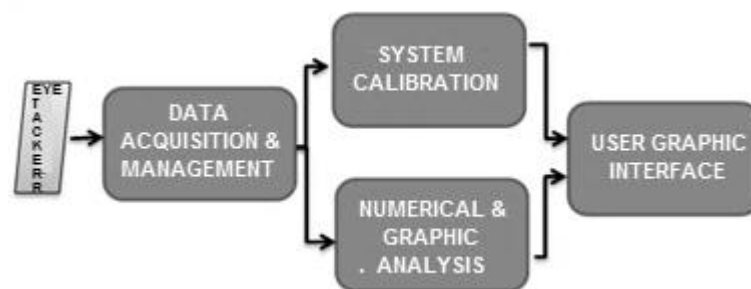


Figure 3.3 Blocks diagram of the proposed eye tracking interface.

3.2.1. Data acquisition and management module

This module allows connecting the eye tracker server and getting data, measured and pre-processed to eliminate noise and reduce erroneous data from the device, in order to facilitate data access (write and read eye tracking data from a text file).

The communication process of the module for data acquisition is done by the functions `DataAcquisition()` and `ConfigurationAcquisition()`, via a sub-server in Python, which receives inputs from Matlab and then sends commands to the Eye Tribe server, because Matlab does not have a suitable multithreading functionality. Figure 3.4 shows the server configuration for the eye tracker data acquisition.

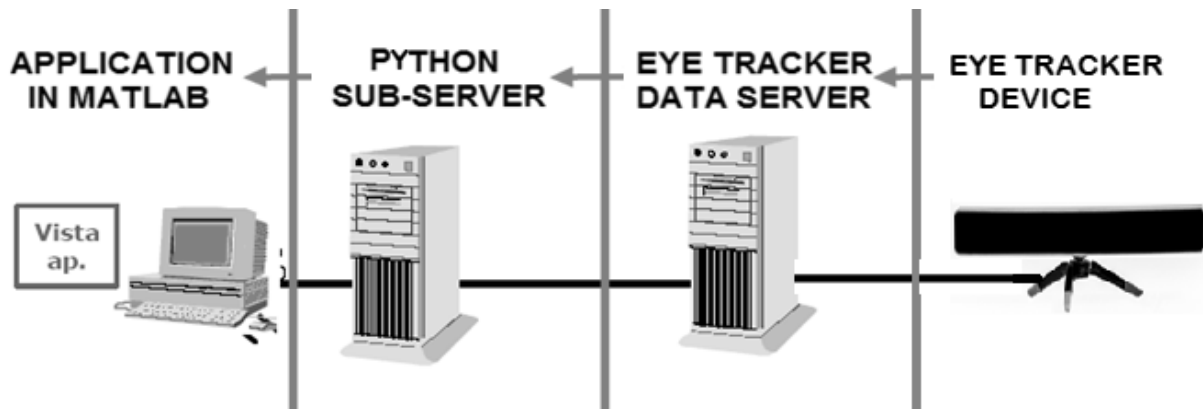


Figure 3.4 Server communication for eye-tracker data acquisition.

This research also proposes a system with a low pass filter `DataFilter()`, to filter the signal acquired from the eye tracker, which helps some users who have difficulty of staring at a fixed point. This filter can be adjusted to both the user's eye speed and eye tracker own latency.

In order to access and manage the eye tracker information, `Save_data_to_file()` and `Read_Data_from_File()` functions to save and read data from a text file, were developed.

3.2.2. Operating set up point calibration and control module

The procedure used for eye tracking set-up, before using it to map the eye gazing onto a screen, is to calibrate the device for each user. Sometimes, there are off-set errors, delays in the velocity of tracking, troubles in selecting the right size of the window and doubts about the appropriate speed of command activation. In order to measure and evaluate those four problems, experimental tests were conducted using the eye-tracker.

In this module, several functions were developed to perform these tests: `Focal_attention()` to calculate the focal point; `Average_of_points_Attention()` to calculate the focal point for a data vector; `Tracking_objects()` to follow trajectories in the screen;

Analysis_trought_Espacial_Windows() and Analysis_trought_Temporal_Window() for analysis in specific sector of the screen or time.

Test 1: Calibration (off-set correction)

Figure 3.5a shows the experimental test to quantify the off-set error. In this figure, five points corresponding to the center, right, left, bottom and top side of the screen (red polygons) are marked. Then the volunteer is asked to look for 5 seconds for each mark in a counter-clockwise sequence. The data are saved and the error is calculated using the equation (3.2) for off-set error estimation, where Ppos is the mark position, ETdata is the estimated position measured with the eye tracker device, and max is the maximum data number.

$$Error = \frac{\sum_{n=1}^{\max} P_{pos} - ET_{data}}{\max} \quad (3.2)$$

Test 2: Velocity of tracking

Figure 3.5b shows the experimental test to estimate the latency of the system. In this experiment, the user must follow a point that changes its position on the screen every 5 seconds. The aim is to measure the time needed for the user to perceive and look to the current point that has changed its position from the previous location on the screen. The user is able to focus the point according to equation (3.3). In this equation, Tlatency is the delay time, which is calculated by subtracting the time the user takes to focus, Tfoco since the screen appeared; num is the number of points to be evaluated.

$$T_{latency} = \frac{\sum_{n=1}^{num} T_{foco} - T_{screen}}{num} \quad (3.3)$$

Test 3: Concentric window size

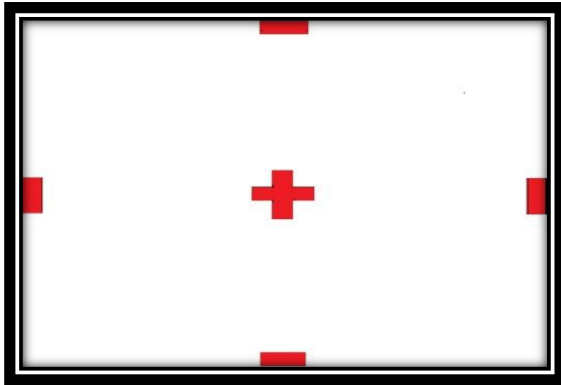
Figure 3.5c shows the experimental test to evaluate different window sizes, which is more appropriated for the function we want to do. In this test, a window with three concentric polygons (150, 100 and 50 pixels size) is displayed for 5 seconds, then the window moves on the screen in 5 different positions. The measured data is evaluated to know what percentage is within each window and know which size would be more appropriate for the user.

Test 4: Command rate

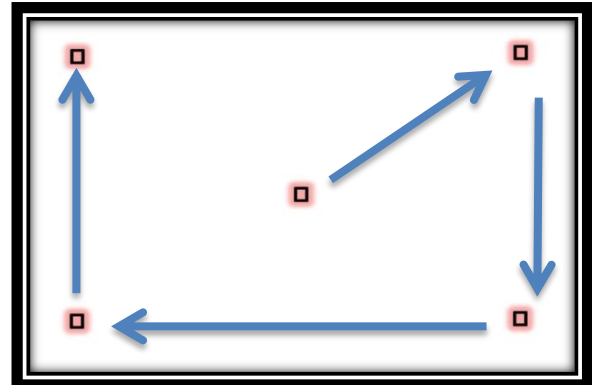
Figure 3.5d shows the experimental test to quantify the average command speed. The system has a graphical application with 9 commands and asks each user to select the commands in ascend order from one to nine. The number of commands per unit of time measures the user

ability to transmit commands and, with this information, it is also possible to calculate the number of errors in those commands.

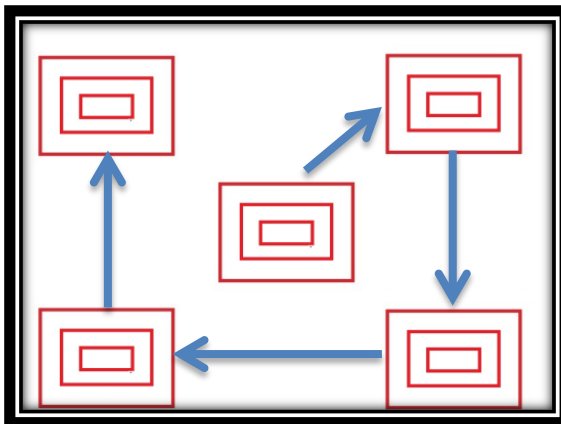
a) Test 1 to estimate the off-set error.



b) Test 2 to estimate the velocity of eye tracking.



c) Test 3 to estimate the concentric windows size.



d) Test 4 to estimate the Command rate.

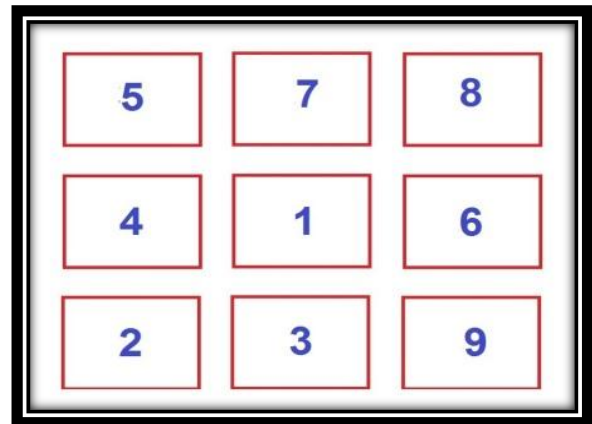


Figure 3.5 Experimental tests to optimize operating set-up point.

3.2.3. Analysis and graphic reports module

The module for analysis and graphic reports is developed to analyze and display different types of graphics using functions, such as: `Time_ET_Graphics()` to plot data of eye tracking and variation in time; `Frequency_ET_Graphics()` to plot the frequency data of eye tracking; `Tracking_in_Image_Video()` to track in an image or a video; `Superposition_ET_images()` to plot superposition with other graphic, image or video; and `Histogram_Graphics()` and `Analysis_Topographic()` to represent data in histograms or topographic images. Figure 3.6 shows examples of this module.

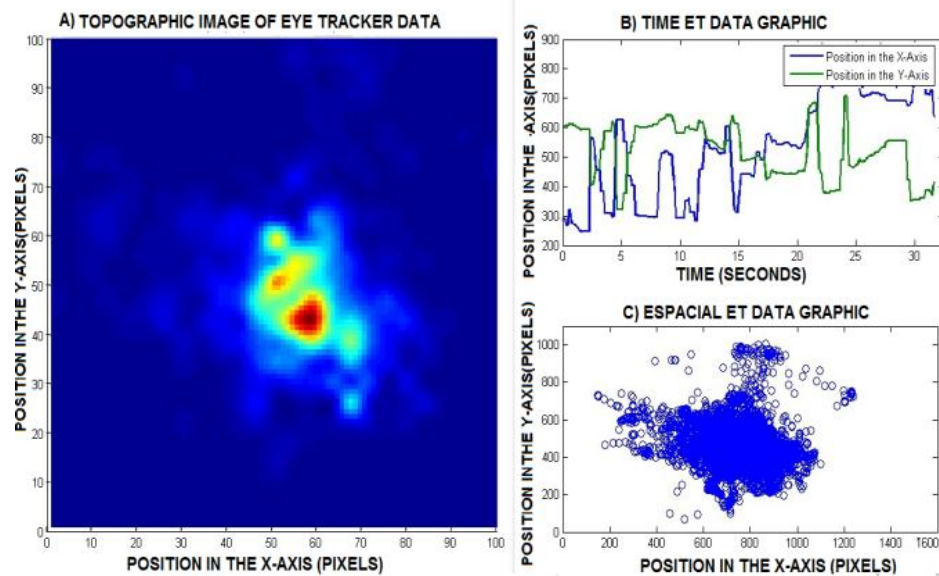


Figure 3.6 Analysis and graphic report. A- Topographic image of eye tracker data, B- Time eye tracking data graphic and C- spatial eye tracking data graphic.

3.2.4. Graphic User Interface (GUI)

The graphic interface developed in this research is a 3 x 3 graphic matrix for a total of 9 commands, a push button to execute the application, another push button for graphic options and a title with information of eye position and activation commands. Figure 3.7 shows the configurable basic user interface.

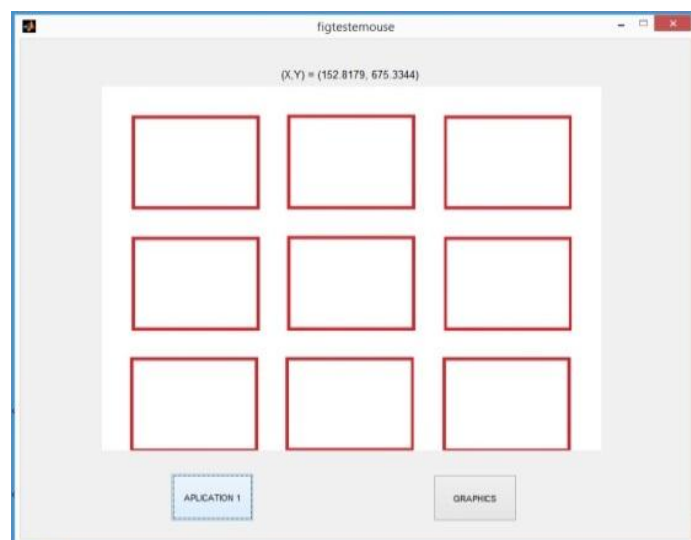


Figure 3.7 Graphic User Interface.

3.3. Analysis and results for the ET ToolBox developed

In this research, an experimental procedure for development of an eye tracking interface (Matlab Toolbox for eye tracking) for assistive applications was developed in order to assess the applicability of an eye tracker device as a tool for visual social attention applications. It was implemented four modules for processing eye tracker data. Figure 3.8 shows the class diagram of the system with the functions and variables for the four modules developed.

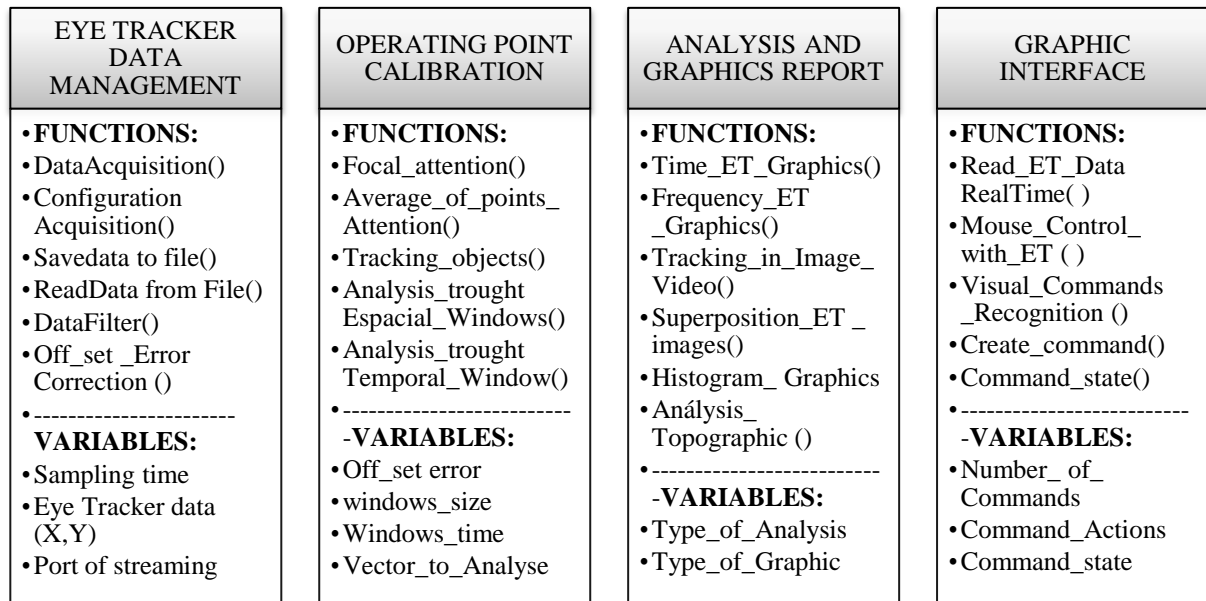


Figure 3.8 Class diagram of the developed eye tracking interface.

This procedure had the participation of sixteen healthy adult volunteers (twelve men and four women), with mean age of 28 years old (± 5.32). Each volunteer was invited to sit comfortably in a chair positioned in front of the screen of a computer (19 inches) and an eye-tracking device, with eyes at 70 cm from screen and at 60 cm from the eye-tracker. Figure 3.9 shows the setup used for the experimental tests. A self-calibration of the eye tracker device was necessary to gather a good data acquisition, which consisted of tracking visually mobile points on the screen and, subsequently, fixed points of known coordinates. The participant viewed a set of fixed and mobile points and windows in a computer screen of 19 inches and 1024 x 768 pixels resolution.

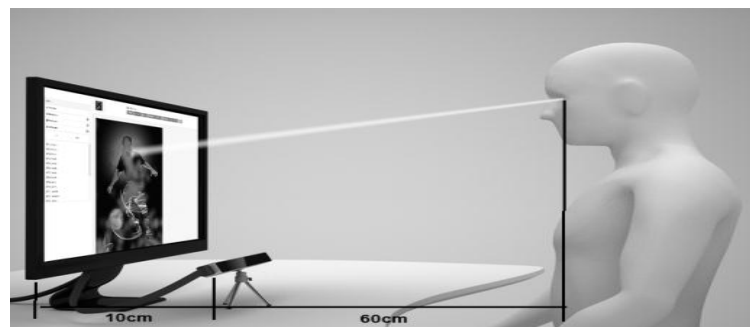


Figure 3.9 Set up for the experimental tests.

Data acquisition and management module results: low pass filter designed to eliminate noise (Butterworth filter) for a maximum velocity of 220ms and with cut frequency of 5Hz. Figure 3.10 shows the output signal (red) for the eye tracker and the blue line for the filtered signal. It is possible to see that the filtering process reduces errors in handling mouse with the eye tracker.

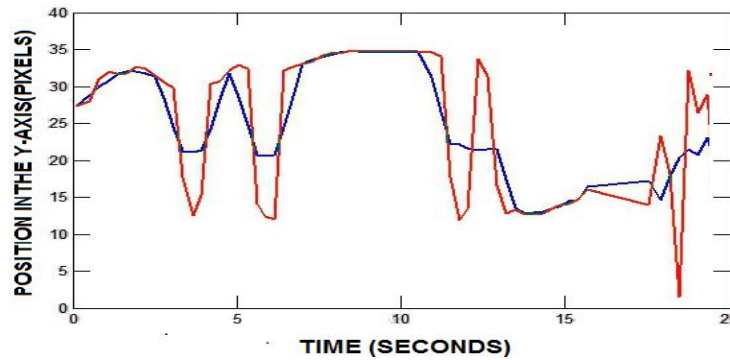


Figure 3.10: Original and filtered signal of the eye tracker, output signal (red) and filtered signal (blue).

Set-up for Point Calibration: Table 3.1 shows the average errors for the off-set calibration in pixels, centimeters and degrees.

Table 3.1: Errors for off-set test.

| Screen axis | Error | Standard deviation |
|-------------|-------|--------------------|
| X(Pixels) | 14.74 | 34 |
| Y(Pixels) | 1.94 | 40 |
| X(cm) | 3.4 | 7.8 |
| Y(cm) | 1.9 | 8.9 |
| X(°) | 0.3 | 0.6 |
| Y(°) | 0.1 | 0.7 |

For the velocity test, the average latency for the system device was of 40 ms and the eye response delay was 192 ms. Table 3.2 shows the results.

Table 3.2: Velocity of tracking test.

| Type of delay | Delay (ms) | Standard deviation |
|-------------------------|------------|--------------------|
| System Latency (ms) | 40 | 33 |
| Eye response delay (ms) | 192 | 45 |

In Tables 3.1 and 3.2 the standard deviation is too high because according to the documentation of the eye tracker device (The Eye Tribe, 2014), the latency of the device is above 16ms and eye response for healthy people is above 200ms.

Table 3.3 shows the test with concentric window size. The result for average data within the window, for windows of 150 pixels was of 95%, for 100 pixels, 85%, and for 50 pixels, 67%.

Table 3.3: concentric window size test.

| Window Size (pixels) | Data within the window (%) | Standard deviation |
|----------------------|----------------------------|--------------------|
| 150 | 95 | 6.3 |
| 100 | 85 | 7.2 |
| 50 | 67 | 8.7 |

Table 3.4 shows the results of the test to estimate the command rate of the system and number of errors for three different command time lengths: 0.5, 1 and 2 seconds.

Table 3.4: Command rate test.

| Command time(seconds) | Command velocity (%) | Number of errors |
|-----------------------|----------------------|------------------|
| 0.5 | 1.5 | 3 |
| 1 | 0.8 | 1 |
| 2 | 0.4 | 0 |

Analysis and graphic reports

Figures 3.6 and 3.10 are examples for the analysis and graphic reports module. All the results in this work were calculated using this module. Another kind of graphic used to analyze the system was the histogram. Figure 3.11 shows an example of validation for six images and four time windows. The histogram calculates the percentage of visual focus time for each image and each time range.

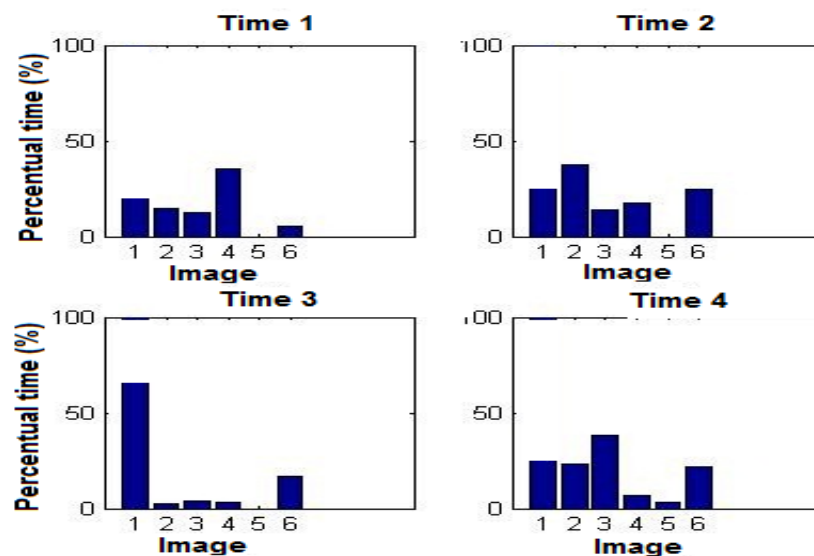


Figure 3.11: Histogram for different images.

Figure 3.12 shows an example of superposition of eye tracker data and image or video files. The system allows real time tracking or off-line superposition in image and video files.

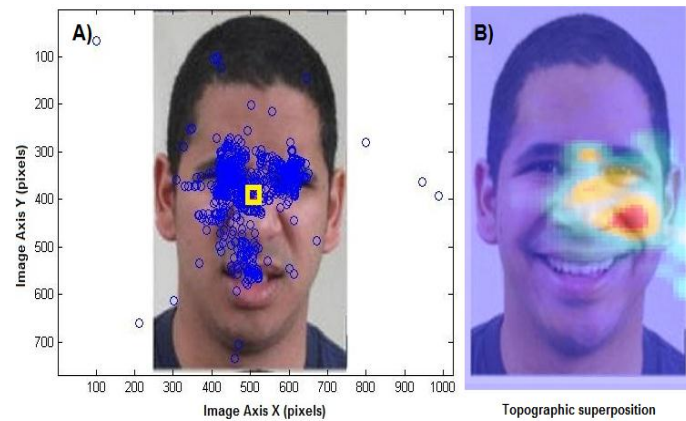


Figure 3.12: Example of superposition A) superposition of data and image; B) superposition of topographic eye tracker and image.

Graphic User Interface (GUI)

The graphic user interface was developed using the results of the different methods to ensure greater reliability and user comfort with the following characteristics: maximum configurable command is 9, but the user can use less commands; the minimal command time of 0.5 seconds, but by default 1 second is recommended for optimum work. The size of each command window is 200 x 150 pixels and the graphic screen is 1024 x 768 pixels to facilitate use in computers, laptops or tablets. The system reliability and the success rate is 90% for 1 second command rate, and 99% for 2 second command rate. Figure 3.13 shows an example of the graphic user interface. The application allows the control of a robot using eye tracking, in which six commands were configured to control four directions, in addition to stop option and a menu.

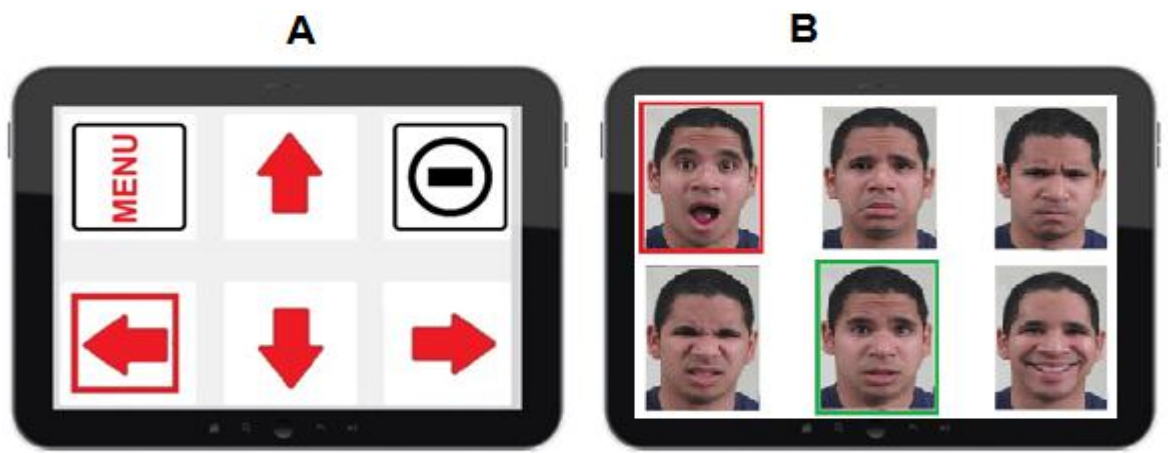


Figure 3.13 GUI Application: A) Robot command using eye tracking, B) Emotion recognition using eye tracking.

3.4. Discussion

The modular toolbox developed in this research allows: acquiring and managing data from the eye tracker, calibrating and preparing the system according to the user disability; analyzing and generating graphical reports; and facilitating the implementation of graphical interfaces controlled by eye gaze using the eye tracker. These modules and their implemented functions are the basis for the eye tracker applications for visual social attention presented in Chapters 6 and 7.

The data acquisition and management module allows connecting Matlab applications with the eye tracker using a Python sub-server. In order to correctly connect them, firstly the eye tracker server should be connected, then the Python sub-server and, finally, the Matlab application should be started.

The configuration and calibration of the system allow testing and setting-up the optimum configuration for each user, according to the disability level, and the motor control of eye movements. This module is independent of the eye tracker device and can be used with other eye tracker systems.

The analysis and graphic report module was developed to allow the study of eye tracker data. The graphics can show the variation in time in the focal point. The frequency analysis can show the average attention focus, the superposition graphics show the part of the screen where the user is viewing, and the histogram analysis is used to compare different regions of interest in the image.

The graphic user interface (GUI) developed is being used to control equipments of an intelligent environment by people with disabilities, motor intention in robotic walkers, studies about valence and emotional based on facial expressions, and in recognition of emotions and focus of attention in children with autism spectrum disorder.

In addition, the use of eye-tracking can benefit other applications, which require observation and evaluation of human attention objectively and non-intrusively, including games, operational system navigation, e-books, market research studies and intelligent environment control.

CHAPTER 4

4. FACIAL EXPRESSION RECOGNITION USING THE KINECT

Automated reading and analysis of human emotion has the potential to be a powerful tool to develop a wide variety of applications, such as human-computer interaction systems, but, at the same time, this is a very difficult issue because the human communication is very complex (Murugappan et Al., 2010). There are different ways of communication, verbal and non-verbal, such as body gestures, speech, facial expressions and hand gestures (Koesltra et Al., 2012). Facial expression communication is especially effective because there are some emotions (called basic emotions), whose expressions are the same over the entire population, in contrast to communication by body gestures, speech or hand gestures, whose elements are different among the cultures throughout the world.

Ekman and Fiesen (1978) developed a Facial Action Coding System (FACS) that describes all possible perceivable facial muscle movements in terms of predefined Action Units (AU). All AUs are numerically coded and facial expressions correspond to one or more AU, based on the FACS system. In this chapter a system to map detected AU to six basic emotions is presented (Figure 4.1). To facilitate the design of this system, five modules are implemented, which allow: data acquisition, face detection, AU features extraction, expression classifier training and expression recognition. The use of Kinect and the methods implemented in this work can benefit automatic emotion recognition applications, which requires observation and evaluation of human expressions objectively and non-intrusively.

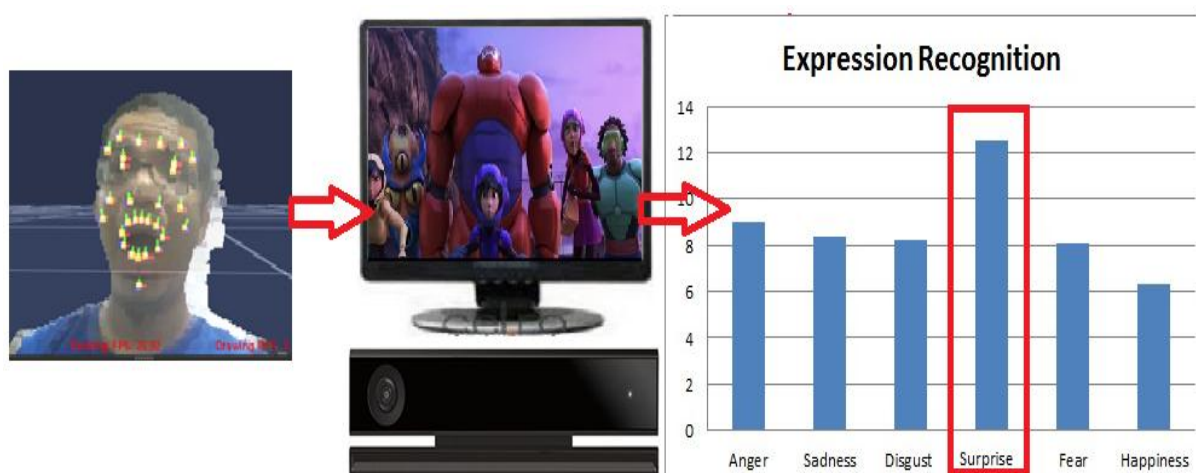


Figure 4.1 Expressions recognition system using Kinect.

4.1 Background: facial expression recognition using FACS and AU system

Facial expressions result from the contraction of facial muscles, making a temporary deformation of the neutral expression. These deformations are typically brief and last mostly between 250ms and 5s according to Fasel and Luetttin (2003). Darwin (1965) is one of the early researchers to explore the evolutionary foundations of facial-expressions display. He argues that facial expressions are universal across humans. He contends that there are habitual movements associated with certain states of the mind. These habits have been favored through natural selection and inherited across generations. Ekman and Fiesen (1978) worked on the idea of facial-expression universality to conceive the facial action coding system (FACS) that describes all possible perceivable facial muscle movements in terms of predefined action units (AUs). All AUs are numerically coded and facial expressions correspond to one or more AUs. Although FACS is primarily employed to detect emotions, it can be used to describe facial muscle activation, regardless of the underlying cause. FACS presented by Ekman and Fiesen (1978) providing a method for objective measurement of facial expressions.

Emotion recognition from facial cues based on FACS rules can be classified as: a) single-phase, where emotions are recognized directly; and b) two-phase, where the facial AU, which are considered as building blocks of facial expressions, are detected first and then the output emotion is inferred from the detected AUs. Then latter approach is found to be more practical than the former, as most of the facial expressions can be described using a sub-set of 44 AUs defined by Paul Ekman. Detecting AUs prior to emotion makes a recognition system more suited to a culture-independent interpretation. Besides, it reduces the amount of independent training data required to model each emotion as there are around 7,000 emotions in practical.

Figure 4.2 shows the six basic expressions described by Paul Ekman, who has also identified a set of facial features. Those features can characterize an expression of each basic emotion.



Figure 4.2 Six basic facial expressions described by Paul Ekman

Sadness: inner corner of eyebrows are raised, eyelids are loose and lip corners are pulled down.

Happiness: muscles around the eyes are tightened, crow's feet wrinkles appears around the eyes, cheeks are raised and lip corners are raised diagonally.

Fear: eyebrows are pulled up and together, upper eyelids are pulled up and mouth are stretched.

Surprise: entire eyebrows are pulled up, eyelids are also pulled up and mouth are widely open.

Anger: eyebrows are pulled down, upper lids are pulled up, lower lids are pulled up and lips may be tightened.

Disgust: eyebrows are pulled down, nose is wrinkled and the upper lip is pulled up.

4.1.1 Facial Action Coding System (FACS)

Ekman and Friesen (1978) developed the FACS for describing facial expressions by Action Units AUs. Of 44 FACS AUs that they defined, 30 AUs are anatomically related to the contractions of specific facial muscles: 12 are from upper face, and 18 are from lower face. AUs can occur either singly or in combination. When AUs occurs in combination, they may be additive, situation in which the combination does not change the appearance of the constituent AU, or non-additive, which is the opposite situation, when the appearance of the constituents does change. Although the number of AU is relatively small, more than 7,000 different AUs combinations have been observed by Scherer and Ekman (1982). FACS provides descriptive power necessary to describe the details of facial expression.

Commonly occurring AUs and some of the additive and non-additive AUs combinations are shown in Tables 4.1 and 4.2. As an example of a non-additive effect, AU 4 appears differently, depending on whether it occurs alone or in combination with AU 1 (as in AU 1+4). When AU 4 occurs alone, the brows are drawn together and lowered. In AU 1+4, the brows are drawn together but are raised due to the action of AU 1. AU 1+2 is another example of non-additive combinations. When AU 2 occurs alone, it not only raises the outer brow, but also often pulls up the inner brow, which results in a very similar appearance to AU 1+2. These effects of the non-additive AUs combinations increase the difficulties of AUs recognition.

Table 4.1 Upper face action units and some combinations (source: Ying-Li, 2001).




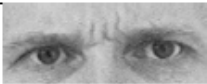






| <i>NEUTRAL</i> | AU 1 | AU 2 | AU 4 | AU 5 |
|---|---|---|--|---|
|  |  |  |  |  |
| Eyes, brow, and cheek are relaxed. | Inner portion of the brows is raised. | Outer portion of the brows is raised. | Brows lowered and drawn together | Upper eyelids are raised. |
| AU 1+2 | AU 1+4 | AU 4+5 | AU 1+2+4 | AU 1+2+5 |
|  |  |  |  |  |
| Inner and outer portions of the brows are raised. | Medial portion of the brows is raised and pulled together. | Brows lowered and drawn together and upper eyelids are raised. | Brows are pulled together and upward. | Brows and upper eyelids are raised. |

Table 4.2 Lower face action units and some combinations (source: Ying-Li, 2001).











| <i>NEUTRAL</i> | AU 9 | AU 10 | AU 20 | AU 26 |
|---|--|--|---|---|
|  |  |  |  |  |
| Lips relaxed and closed. | The infraorbital triangle and center of the upper lip are pulled upwards. Nasal root wrinkling is present. | The infraorbital triangle is pushed upwards. Upper lip is raised. Causes angular bend in shape of upper lip. Nasal root wrinkle is absent. | The lips and the lower portion of the nasolabial furrow are pulled pulled back laterally. The mouth is elongated. | Lips are relaxed and parted; mandible is lowered. |
| AU9+25 | AU9+17+23+24 | AU10+17 | AU 10+25 | AU 9+17 |
|  |  |  |  |  |

Table 4.3 lists the names, numbers and anatomical basis of each AU. Most of the AU involves a single muscle. The numbers are arbitrary and do not have any significance except that 1 through 7 refer to brows, forehead or eyelids. The table indicates where more than one muscle collapses into a single AU, or where distinguished AU are represented by a single muscle. The FACS names given in the table are a shorthand, not meant to describe the appearance changes, but a convenience to call them to mind.

Table 4.3 Action Units list in FACS system (Source: Ekman 1982).

| AU Number | FACS Name | Muscular Basis |
|-----------|----------------------------|---|
| 1 | Inner Brow Raiser | Frontalis, Pars Medialis |
| 2 | Outer Brow Raiser | Frontalis, Pars Lateralis |
| 4 | Brow Lowerer | Depressor Glabellae; Depressor Supercilli; Corrugator |
| 5 | Upper Lid Raiser | Levator Palpebrae Superioris |
| 6 | Cheek Raiser | Orbicularis Oculi, Pars Orbitalis |
| 7 | Lid Tightener | Orbicularis Oculi, Pars Palpebralis |
| 8 | Lips Toward Each Other | Orbicularis Oris |
| 9 | Nose Wrinkler | Levator Labii Superioris, Alaeque Nasi |
| 10 | Upper Lip Raiser | Levator Labii Superioris, Caput Infraorbitalis |
| 11 | Nasolabial Furrow Deepener | Zygomatic Minor |
| 12 | Lip Corner Puller | Zygomatic Major |
| 13 | Cheek Puffer | Caninus |
| 14 | Dimpler | Buccinator |
| 15 | Lip Corner Depressor | Triangularis |
| 16 | Lower Lip Depressor | Depressor Labii |
| 17 | Chin Raiser | Mentalis |
| 18 | Lip Puckerer | Incisivii Labii Superioris; Incisivii Labii Inferioris |
| 20 | Lip Stretcher | Risorius |
| 22 | Lip Funneler | Orbicularis Oris |
| 23 | Lip Tightner | Orbicularis Oris |
| 24 | Lip Pressor | Orbicularis Oris |
| 25 | Lips Part | Depressor Labii, or Relaxation of Mentalis or Orbicularis |
| 26 | Jaw Drop | Masseter; Temporal and Internal Pterygoid Relaxed |
| 27 | Mouth Stretch | Pterygoids; Digastric |
| 28 | Lip Suck | Orbicularis Oris |
| 38 | Nostril Dilator | Nasalis, Pars Alaris |
| 39 | Nostril Compressor | Nasalis, Pars Transversa and Depressor Septi Nasi |
| 41 | Lid Droop | Relaxation of Levator Palpebrae Superioris |
| 42 | Slit | Orbicularis Oculi |
| 43 | Eyes Closed | Relaxation of Levator Palpebrae Superioris |
| 44 | Squint | Orbicularis Oculi, Pars Palpebralis |
| 45 | Blink | Relaxation of Levator Palpebrae and Contraction of Orbicularis Pars Palpebralis |
| 46 | Wink | Orbicularis Oculi |

4.1.2 Automatic facial features extraction and AU recognition

Automatic recognition of FACS-AU is a difficult problem and relatively few works have been reported. AUs have no quantitative definitions and, as noted, they can appear in complex combinations. Mase (1991) and Essa (1997) described patterns of optical flow that corresponded to several AUs, but did not attempt to recognize them. Bartlett et al. (1999) and Donato et al. (1999) reported some of the most extensive experimental results of upper and

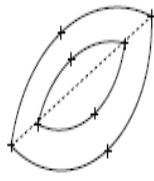
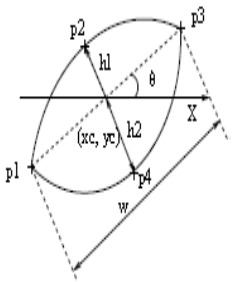
lower face AUs recognition. An automated facial expression analysis system must solve two problems: facial feature extraction and facial expression classification.

4.1.3 Facial feature extraction

Contraction of the facial muscles produces changes in the direction and magnitude of the motion on the skin surface and in the appearance of permanent and transient facial features. Examples of permanent features are the lips, eyes, and any furrows that have become permanent with age. Transient features include facial lines and furrows that are not present at rest but appear with facial expressions. Even in a frontal face, the appearance and location of the facial features can change dramatically.

Table 4.4 shows an example for lips feature extraction: a state lip model represents open, and closed. A different lip contour template is prepared for each lip state. The open and closed lip contours are modeled by two parabolic arcs, which are described by six parameters: the lip center position (x_c, y_c), the lip shape (h_1, h_2 and w), and the lip orientation (μ). For tightly closed lips, the dark mouth line connecting the lip corners represents the position, orientation, and shape. For the eyes, brows, cheeks, furrows, etc, it is possible to obtain different state model representations and, then, extract their features.

Table 4.4. Multi-state facial component models of a lip (source: Ying-Li, 2001)

| Component | State | Description/Feature |
|-----------|--------|---|
| Lip | Open |  |
| | Closed |  |

4.1.4 Facial expression classification

Since each AU is associated with a specific set of facial muscles, using an accurate geometrical modeling and tracking of facial features will lead to better recognition results. Furthermore, the knowledge of exact facial feature positions could be useful for the area-based (Yacoob and Davis, 1996), holistic analysis (Bartlett et al., 1999), and optical flow based (Lien et al., 2000) classifiers.

Figure 4.3 depicts the overall structure of the Automatic Facial Action Analysis AFA system. Given an image sequence, the region of the face and approximate location of the individual's face features are detected automatically in the initial frame (Rowley, 1998). Furthermore, the contours of the face features and components are adjusted in the initial frame. Both permanent (e.g., brows, eyes, lips) and transient (lines and furrows) face features changes are automatically detected and tracked in the image sequence. Informed by FACS AUs, the facial features can be grouped into separate collections of feature parameters, since the facial actions in the upper and lower face are relatively independent for the AUs recognition (Ekman and Friesen, 1978). In the upper face, 15 parameters describe shape, motion, eye state, motion of brow and cheek, and furrows. In the lower face, 9 parameters describe shape, motion, lip state, and furrows. These parameters are geometrically normalized to compensate image scale and in-plane head motion.

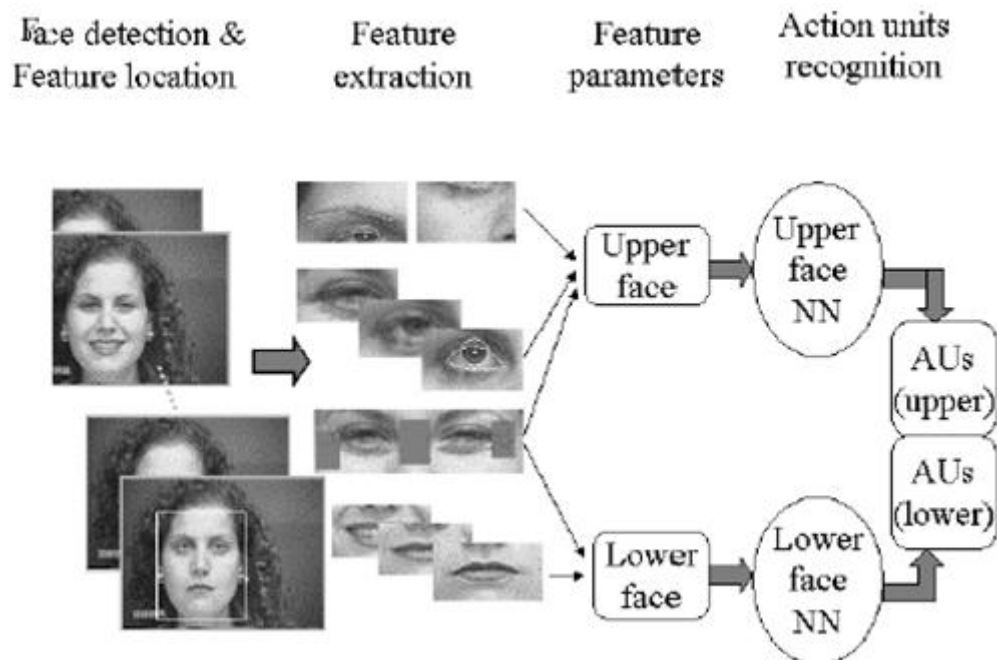


Figure 4.3. Feature-based Automatic Facial Action Analysis (AFA) system (source: Ying-Li, 2001)

4.2 Implementation of the system for expression recognition

Figure 4.4 shows the block diagram of the method for expression recognition developed in this research based on the FACS-AU system. To facilitate the design of this interface, five modules were implemented, which consist on: data acquisition module for acquiring and managing data from the Kinect; face detection module for detecting the face and FACS points; AU features extraction module, to allow facial AU recognition; and a train classifier and expression recognition modules, for expression detection and classification.

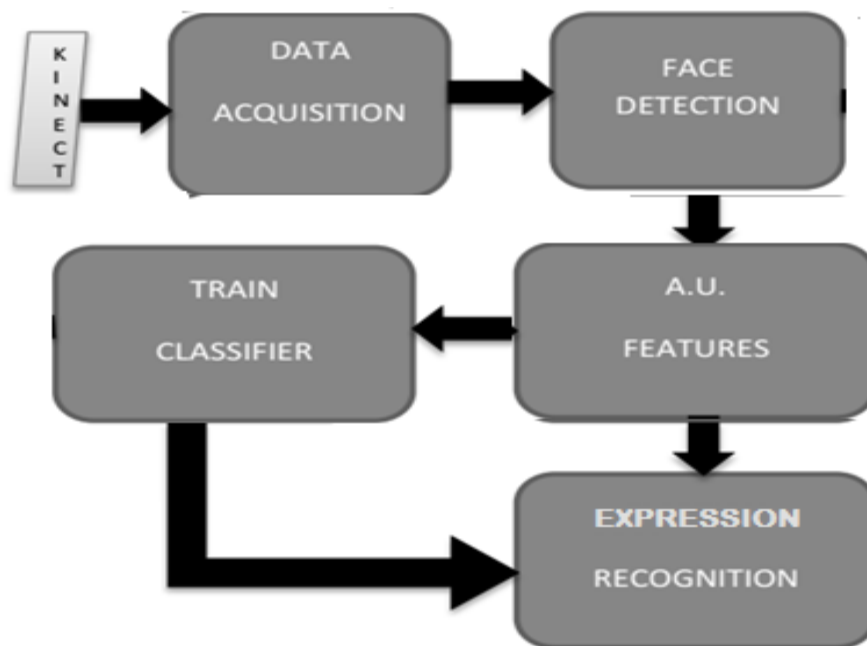


Figure 4.4 Block diagram of the proposed Expression recognition system

4.2.1 Data acquisition

The data acquisition module is composed of one Microsoft Kinect 2.0 device, which provides high quality color, infrared and depth images (Figure 4.5) that are used to obtain a 3D-positions facial model. The algorithms for data acquisition, filtering and preprocessing were developed by methods from Microsoft.Toolkit.FaceTracking (Kinect for Windows SDK, 2013).



Figure 4.5 Kinect data acquisition: A) Depth image; B) Infra-red image; C) Color image 4.5
Kinect data acquisition: A) Depth image; B) Infra-red image; C) Color image.

4.2.2 Face feature extraction: Action Units (AUs)

Functions provided by the Brekel proFace 2 Software applications were used for face detection and AU features extractions.

Face detection module

Automatic functions provided by Brekel proFace 2 Software are used to detect a 3D facial model, which is based on Colombo (2006) and Nair (2009) methods and using curvature features to detect high curvature areas, such as the nose tip and eye cavities. Segmentation is also applied to 3D face detection. Once the face is detected, geometric correspondence between the captured geometry and a model is found. For this the Iterative Closest Point (ICP) is used iteratively to align the closest points between the two shapes in the same method shown by Alyuz (2012). In such method, visible patches of the face are detected and used to discard obstructions before using ICP for alignment. This technique allows the matched 3D model to deform. In Mao (2004), a correspondence is established between landmarks of the model and the captured data face, using a model to deform the shape to match 3D points to the FACS system model. Figure 4.6 shows an example of 3D points detection, extraction and match to 3D model FACS using the Brekel proFace 2.



Figure 4.6 Face detection and 3D facial model creation.

Action Units Features

The AUs features module allows obtain 20 AU features, which are the value of 20 action units. Those values are represented as a set of values that range between 1 and -1, which can be treated as a vector from a 20-dimensional space. Figure 4.7 shows the module for AU feature extraction.

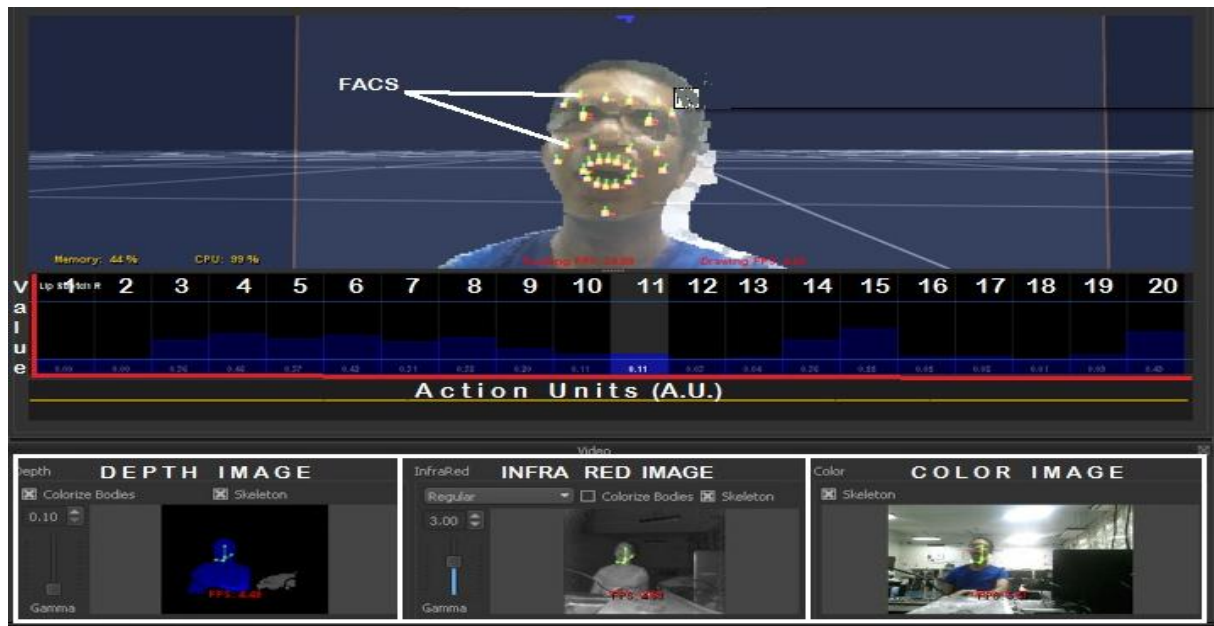


Figure 4.7 Module for AUs feature extraction.

Table 4.5 describes the AUs extracted by the features detection module. This module allows obtaining 9 bilateral (left/right face side) and 2 unilateral AUs, thus corresponding to 20 action units.

Table 4.5. Description of the 20 AUs features detected in this system.

| N° | AU description | Quantity |
|-------|------------------------------|-----------------|
| 1-2 | AU(1) InnerBrowRaiser | 2 (Left/Right) |
| 3-4 | AU(2) OuterBrowRaiser | 2 (Left/Right) |
| 5-6 | AU(4) BrowLowerer | 2 (Left/Right) |
| 7 | AU(10) UpperLipRaiser | 1 (Unilateral) |
| 8-9 | AU(12) LipCornerPuller | 2 (Left/Right) |
| 10-11 | AU(13) CheekPuffer | 2 (Left/Right) |
| 12-13 | AU(20) LipStretcher | 2 (Left/Right) |
| 14-15 | AU(15) LipCornerDepressor | 2 (Left/Right) |
| 16 | AU(26) JawLowerer | 1 (Unilateral) |
| 17-18 | AU(43) EyesClosed | 2 (Left/Right) |
| 19-20 | AU(47) JawLeftRight | 2 (Left/Right) |

4.2.3 Expression recognition

In the emotion detector module, in order to classify the user emotion based on 20 AU, K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) algorithms were used, based on the work of Jiawei et al. (2012).

K-Nearest Neighbors (KNN) algorithm

The KNN algorithm allows to predict a value of variables (20-dimensional AU vector), and classifying them into different classes (six basic facial expressions). The main assumption of this algorithm is that the state of the observed AU vector can be classified based on previous observations of similar AU vectors, which were classified with the same features. The AU vector is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its KNN, where K is a positive number, equal or greater than 1.

KNN algorithm is a type of fuzzy learning classification method. It means that all computations of similar objects (AU vectors) from training examples to the object that is currently being classified are made during the classification process. The training set was always created by learning process. In this case, this process is based on all previous observations. A very interesting issue in this algorithm is how to find a similarity between two objects based on their features. In Jiawei et al. (2012), KNN emotion detector is used to compute the emotion using Euclidean distance, because, as said previously, the emotion features (which are the value of 20 AUs) are represented as set of value between 1 and -1. The basic Euclidean distance for the two-dimensional space is represented in Equation (4.1):

$$D(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (4.1)$$

However in this work the algorithm, which should find distance between 20-dimension vectors, the formula used is represented by Equation (4.2):

$$D(p_1(x_1, \dots, x_n), \dots, p_n(y_1, \dots, y_n)) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2} \quad (4.2)$$

Where: D is the distance (similarity); p_n is the object; x_1, \dots, x_n and y_1, \dots, y_n are the features sets.

By using it, the algorithm can find a similarity of each object from the training set to currently classified object and choose the K with most similar objects. In classification mechanism, it also has been used a normalization formula for all results after computing the similarity among the objects.

Normalization is a process to adjust values which are measured on different scale to strictly specified range. Often it is made in order to allow easier data comparison. In this research, a Min – Max normalization method was used, which is based on the work of by Jiawei et al. (2012).

Linear Discriminant Analysis (LDA) algorithm

LDA or Fisherfaces method (Belhumeur et al., 1997) overcomes the limitations of the eigenfaces method by applying the Fisher's linear discriminant criterion. This criterion tries to maximize the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples.

The LDA method tries to find the subspace that best discriminates different facial expressions classes. The within-class scatter matrix, also called intra-personal, represents variations in appearance of the same individual due to different lighting and face expression, while the between-class scatter matrix, also called the extra-personal, represents variations in classes.

In this research, one maximize the distance between the face AU of different classes. One minimize the distance between the face AU of the same class. In other words, the objective is to maximize the between-class scatter (SB), while minimizing the within-class scatter matrix (SW) in the projective subspace.

The within-class scatter matrix (SW) and the between-class scatter matrix (SB) are defined as in Equation (4.3).

$$S_W = \sum_{j=1}^C \sum_{i=1}^{N_j} (\Gamma_i^j - \mu_j) (\Gamma_i^j - \mu_j)^T \quad (4.3)$$

where Γ_i^j is the i th sample of class j , μ_j is the mean of class j , C is the number of classes, N_j is the number of samples in class j . In Equation (4.4), it is defined how the scatter matrix (SB) is calculated:

$$S_B = \sum_{j=1}^C (\mu_j - \mu) (\mu_j - \mu)^T \quad (4.4)$$

where μ represents the mean of all classes. The subspace for LDA is spanned by a set of discriminant vectors $W = [W1, W2, \dots, Wd]$, satisfying Equation (4.5):

$$W = \arg \max = \left| \frac{W^T S_B W}{W^T S_W W} \right| \quad (4.5)$$

The within-class scatter matrix expresses how closely facial AU are distributed within the classes, while the between-class scatter matrix quantifies how separated the classes are from each other. When face AUs are projected onto the discriminant vectors W , facial AUs should be distributed closely within the classes and separately between the classes, as much as possible. In other words, these discriminant vectors minimize the denominator and maximize the numerator in Equation (4.5). Therefore, W can be constructed with the aid of the eigenvectors of $S_W^{-1} S_B$.

4.3 Analysis and results

Procedure

In this test, each volunteer is asked to sit comfortably in a chair positioned in front of a 19-inches computer screen and a Kinect sensor, with his/her eyes at 60 cm away from the screen and at 50 cm from the sensor. The screen displayed the six pictures relative to human facial expressions (Figure 4.9) for ten seconds. The participant should imitate each emotional expression three times, as shown in Figure 4.8.

The Kinect device recorded images of each emotional facial expression performed by the volunteer, and our algorithm based on KNN or LDA identified the set of features related to expressions of each emotion, based on the AUs. The test involved the participation of eight healthy adults, aged between 24 and 33 years (M: 26, SD: ± 3.81).



Figure 4.8: Experimental procedure. Participants imitating the model of emotion facial expression displayed on the screen.

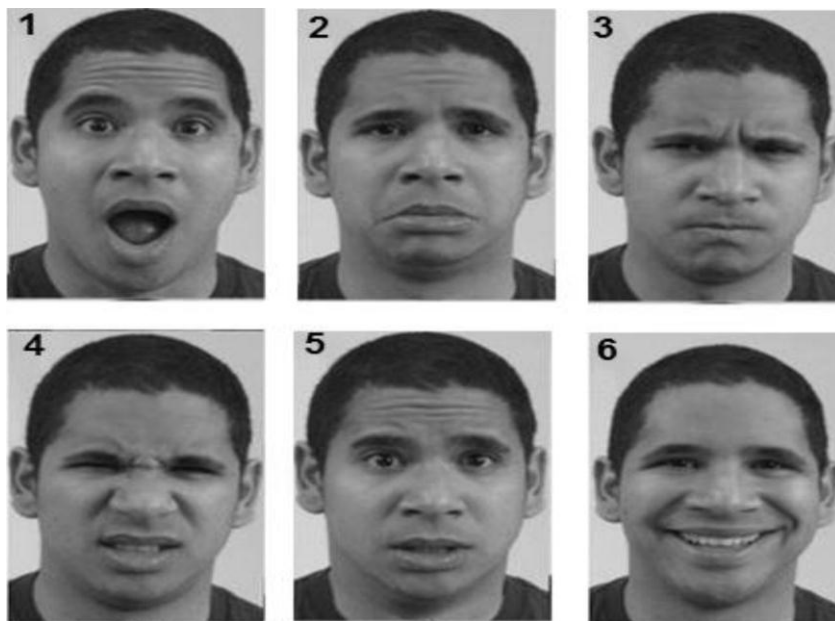


Figure 4.9: Emotional facial expressions viewed by the participants. E1 (surprise), E2 (sadness), E3 (anger), E4 (disgust), E5 (fear) and E6 (happiness) (Source: Du 2014).

AU features

Figure 4.10 shows the 20 AUs signals from the eight volunteers imitating ten seconds the six basic expressions, these vectors are used for our algorithm based on KNN or LDA to identified the set of expressions of each emotion

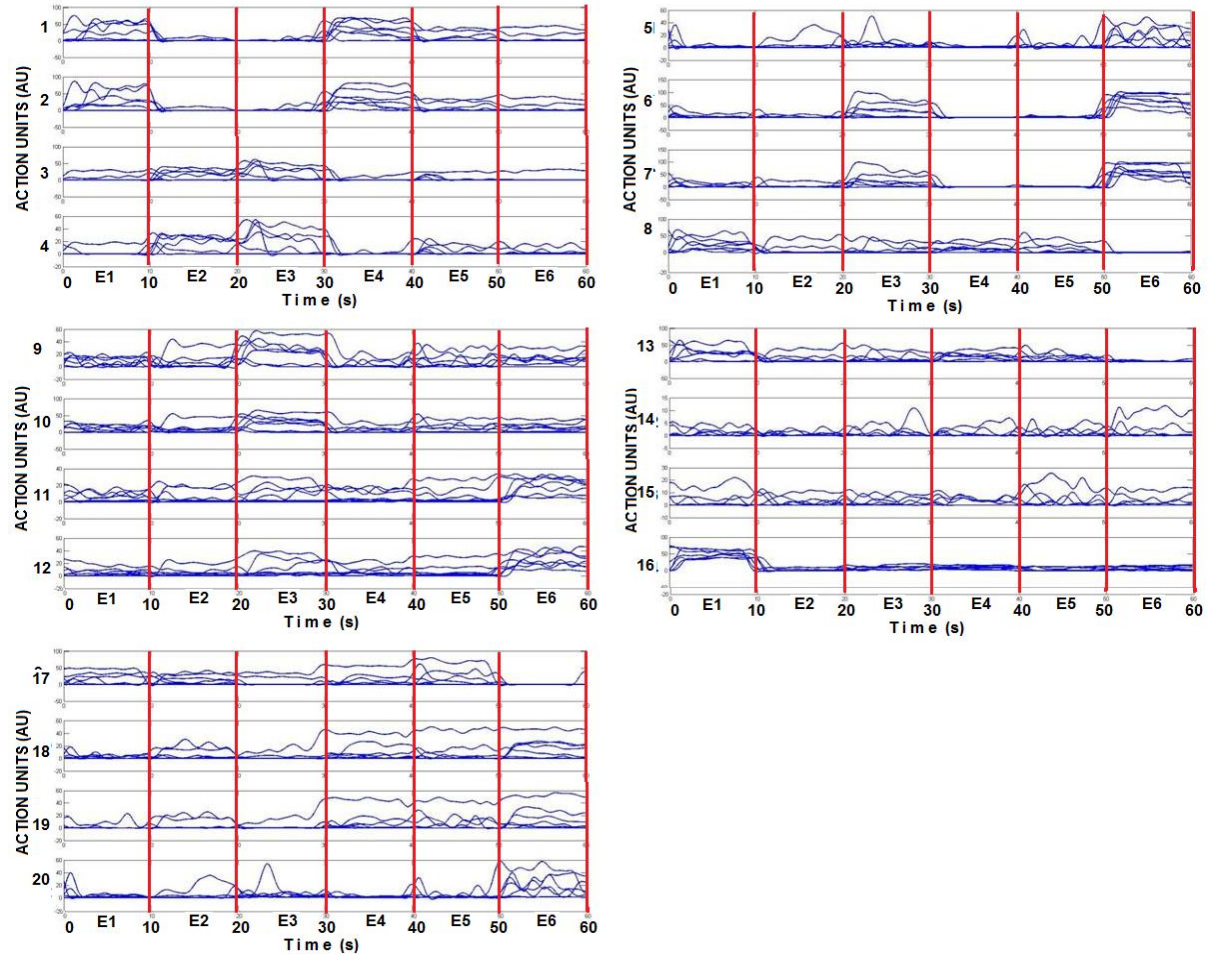


Figure 4.10 Twenty AUs signals obtained from eight different volunteers imitating the six basic expressions

Regarding the identification of facial emotions by the Kinect device, using the LDA classifier for the three emotional classes (positive, negative and neutral), they were recognized with an overall accuracy of 80.1%, being 83.6% for neutral, 81.1% for negative class and 75.8% for positive. In turn, the KNN classifier recognized an overall accuracy of 82.9%, being 87.5% for neutral, 84.2% for negative class and 77.7% for positive, as shown in Table 4.6. Thus, the recognition system of emotional classes was able to identify a large quantity of specific facial features related to the neutral emotional class.

Table 4.6 Accuracy of the emotion recognition for three class.

| Emotion recognition | LDA Accuracy | KNN Accuracy |
|---------------------|--------------|--------------|
| Positive Class | 75.8% | 77.7% |
| Negative Class | 81.1% | 84.2% |
| Neutral Class | 83.6% | 87.5% |
| Total | 80.1% | 82.9% |

Expressions recognition

Regarding the identification of facial emotions by the Kinect device and considering the six emotional classes (anger, fear, sadness, happiness, surprise, disgust), the LDA classifier identifying them with an overall accuracy of 62%, while the KNN classifier recognized the classes with an overall accuracy of 70%, as shown in Table 4.7 and 4.8. Thus, the recognition system of emotional classes was able to identify a large quantity of specific facial features.

Table 4.7 Confusion matrix for six emotion recognition using LDA

| Expression | Anger | Fear | Sad | Happiness | Surprise | Disgust |
|------------|-----------|-----------|-----------|-----------|-----------|-----------|
| Anger | 49 | 12 | 15 | 15 | 2 | 17 |
| Fear | 6 | 55 | 12 | 1 | 12 | 2 |
| Sadness | 3 | 4 | 48 | 3 | 1 | 0 |
| Happiness | 24 | 3 | 7 | 70 | 5 | 9 |
| Surprise | 7 | 25 | 6 | 5 | 80 | 0 |
| Disgust | 11 | 1 | 12 | 6 | 0 | 72 |

Table 4.8 Confusion matrix for six emotion recognition using KNN

| Expression | Anger | Fear | Sad | Happiness | Surprise | Disgust |
|------------|-----------|-----------|-----------|-----------|-----------|-----------|
| Anger | 58 | 6 | 11 | 12 | 3 | 11 |
| Fear | 7 | 64 | 15 | 0 | 5 | 3 |
| Sadness | 2 | 5 | 50 | 4 | 0 | 2 |
| Happiness | 21 | 4 | 6 | 89 | 3 | 5 |
| Surprise | 8 | 20 | 8 | 3 | 85 | 4 |
| Disgust | 4 | 1 | 10 | 2 | 4 | 75 |

4.4 Discussion

In this chapter, was developed a system for expression recognition based on the FACS-AU system to classify six basic expressions. The developed system allows: image acquisition (IR, depth and color), face detection (FACS 3D model), features extraction (20-dimensional AU vector), classifiers training (LDA or KNN) for six expression recognition (anger, fear, sadness, happiness, surprise, disgust and neutral). This system and the implemented functions are the basis for the emotion recognition applications, which are presented in Chapters 6 and 7.

The image acquisition module allows connecting the Brekel application with the Kinect device using the Microsoft Toolkit FaceTracking library from Kinect for Windows SDK 2. Then, a 3D facial model is obtained and 20 AU features are detected using the face detection

and feature extraction module of Brekel. The quality of the AUs depends on the illumination, head movements, distance to the sensor, and facial characteristics and accessories (hair length, bangs, beard, mustache, glasses)

The expression recognition module is based on trained classifiers (LDA or KNN). Facial expressions for three classes (negative, neutral and positive) were recognized by the computational system, with accuracy rates of 80.1% and 82.9% for LDA and KNN classifiers, respectively. For six classes, the accuracy rates for LDA was 62.3% and for KNN, it was 70.1%.

CHAPTER 5

5. EMOTION DETECTION USING THERMAL CAMERA

Emotions are often perceived in the body and face, where it becomes apparent that changes in physiological conditions arise from emotional states. The temperature is a kind of, being a physiological indicator used as psychological marker of emotions. Studies suggest that Infrared Thermal Imaging (IRTI) may assist detection, recognition, and tracking of faces, classification of facial expressions, and Automated Affect Interpretation (AAI). Contraction or expansion of facial muscles causes fluctuations in the rate of blood flow. Noninvasive detection of any changes in facial thermal features may help in detecting, extracting, and interpreting facial expressions or emotions. However, a representative model for estimating the relationship between fluctuations in blood flow and facial emotional activity is not yet available.

The main goal of the present chapter is to ascertain whether facial thermograms can be used as a valid and reliable somatic indicator of emotional parameters. Specifically, this work wants to determine if there is a relation between changes in facial temperature and valence, arousal and subjective feelings. In this chapter, a system for studying the use of thermography as an experimental paradigm to recognize emotions and discover the relationship between facial thermal variation and emotional activity is presented. To facilitate the design of this system, four modules are implemented, which allow: data and thermal images acquisition; Facial Thermal - Region of Interest (FT-RoI) segmentation; features extraction; and detection of facial thermal variation. Figure 5.1 shows a diagram of the system.

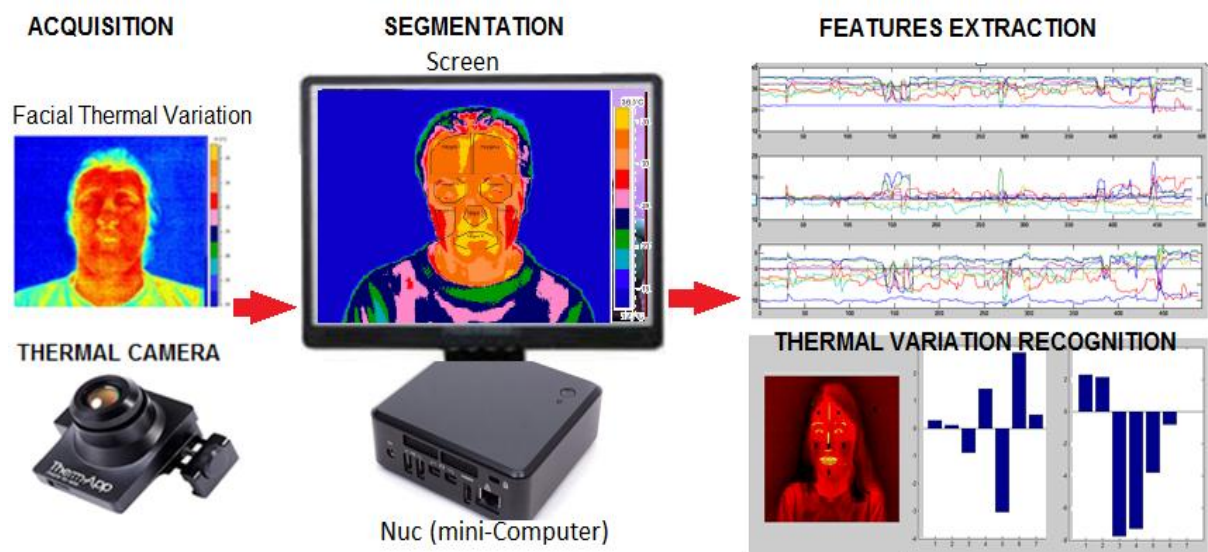


Figure 5.1 System used to study facial thermal variation detection.

5.1. Background: application of thermography to study of emotions

Few studies have applied thermography to study emotions. Pavlidis, Eberhardt, and Levine (2002) have used infrared cameras to measure participants' facial temperatures, based on the idea that facial temperature changes in various regions of the face correlate with emotional experience. In particular, they have studied the emotions of deceit and anxiety and found evidence that facial temperature changes can indeed predict both. However, some results of thermographic studies are sometimes inconclusive.

Briese and Cabanac (1991) found that stress levels correlate with increased blood flow in the frontal vessels of the forehead. On the other hand, Tanaka, Ide and Nagashima (1999), and Nagumo, Zenju, Nozawa, Ide and Tanaka (2002) obtained correlations between arousal level and nasal skin temperature. Zenju, Nozawa, Tanaka, and Ide (2004) found that nasal skin temperature increases when shifting to pleasant mental states and decreases when shifting to unpleasant mental states. Similarly, Kuraoka and Nakamura (2011) obtained decreased nasal temperature in negative emotional states, but Nakanishi and Imai-Matsumura (2008) observed facial skin temperature decrements also during joyful expressions in the nose. The correlations between facial thermal changes and other brain or physiological measures are clearly significant during experimental tasks. However, researchers such as Khan, Ward, and Ingleby (2006, 2009) have opened new lines of research in this area: the relation between thermographic changes and feelings. Their experiments show variations in the intensity of the temperature in subjects that express positive and negative affective states, particularly in states of happiness and sadness. In short, thermography can be considered a biometric measurement of human emotions, but arousal, valence, basic emotional states, stress, empathy or feelings, including complex emotions such as love or happiness, are not differentiated in previous research. The characteristics of the populations employed (adults, infants, elderly or animals), the ecological or laboratory context, and the different tests and stimuli employed yield contradictory results, such as the thermal increments or decrements associated with empathy or positive emotions.

The techniques most used for the analysis of facial variation are those based on general regions of interest: nose, mouth, forehead, etc. (named FT-RoI), and those based on more specific points of facial AUs or muscles (Facial Thermal Feature Points – FTFP). Figure 5.2 shows an example of FT-RoI and FTFP.

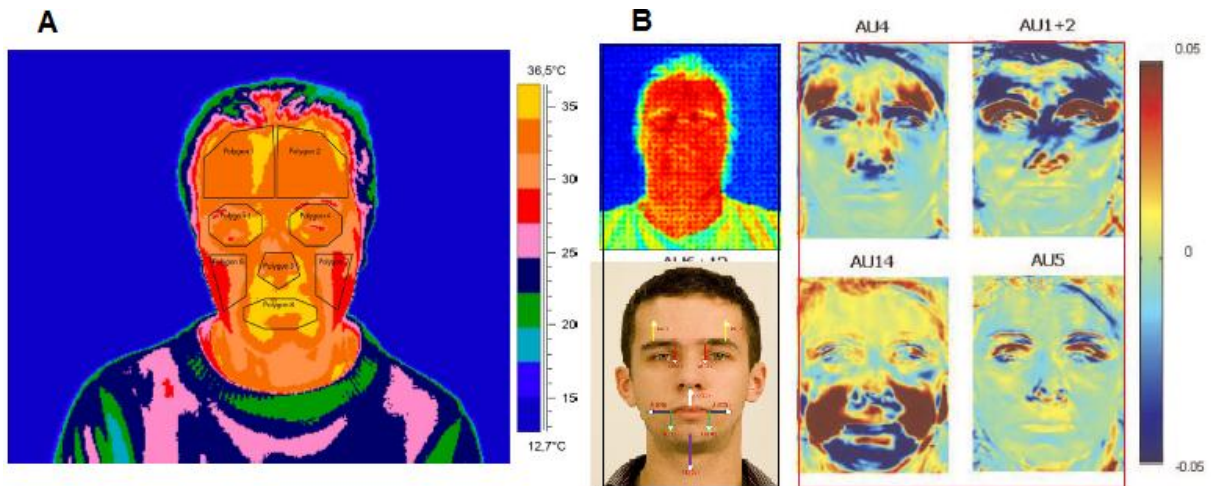


Figure 5.2 Techniques for thermal variation detection. A) Facial Thermal - Region of Interest (FT-RoI); B) Facial Thermal Feature Points (FTFP). (Source: Salazar-López 2015)

5.1.1. Facial Thermal - Region of Interest (FT-RoI)

FT-RoI allows measuring skin temperature variation around certain parts of the face for examining the autonomic nervous activity, such as shown in Figure 5.3. The autonomic nervous system's response to stress or emotional causes a change in the temperature of the skin, which the experimenters measure in the nose, a part of the body that, despite experiencing little movement, can undergo variations in temperature under stressful or emotional conditions. Results of researches reveal a decrease in nasal temperature during stressful situations due to vasoconstriction, which leads to a reduction of blood flow to the peripheral capillaries of the nose, causing the decrease in temperature. Veltman and Vos (2005) claim that the change in nasal temperature is an important measurement, but not the absolute value of the temperature (considering that mental workload may not be the only factor that affects nose temperature). In their study, they used thermographic cameras and determined as Region of Interest (RoI) the nose and forehead, as forehead is one of the most stable temperatures in the body. Their paradigm confirmed the equivalency of temperatures of forehead and nose in rest condition, and checked the temperature changes in the nose in all of the conditions in which mental workload was used. Khan, Ward, and Ingleby (2009) also studied the relations between thermographic changes and feelings using RoIs. Their experiments show variations in the intensity of the temperature in subjects that express positive and negative valence states, particularly in states of happiness and sadness.

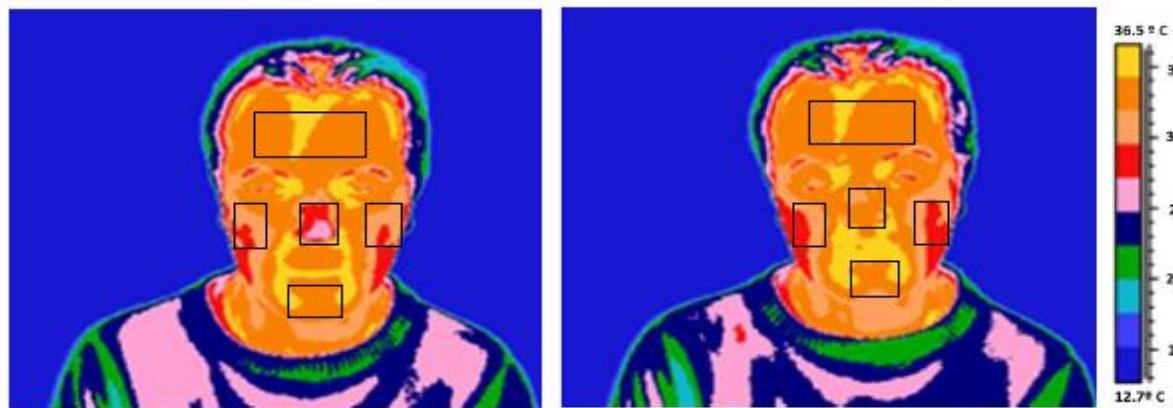


Figure 5.3 Example of Facial Thermal – Region of Interest (FT-ROI).

5.1.2. Facial Thermal Feature Points (FTFP)

FTFP is a noninvasive technique for Automated Facial Expression Classification (AFEC) and Automated Affect Interpretation (AAI). Recent studies suggest that FTFP may assist detection, recognition and tracking of faces, classification of facial expressions, and AAI (Eveland et al. 2003; Sugimoto et al. 2000). Contraction or expansion of the facial muscles (thermo-muscular activity) causes fluctuations in the rate of blood flow, which results in a change in the volume of blood flow under the surface of the facial skin. Infrared Thermal Imaging (IRTI) can help detecting the change in blood flow volume, thus following thermo-muscular activity through skin temperature measurements.

In fact, contactless detection of any changes in facial thermal features may help detecting, extracting, and interpreting facial expressions. However, a representative model for estimating the relation between fluctuations in blood flow and facial muscle activity is not yet available. Such a model could enhance the understanding of the relation between facial expressions and the facial thermal according to physiological characteristics.

A small number of attempts to analyze facial expressions using IRTI, singly or in combination with other cues, have been tried (Khan et al. 2005; Pavlidis 2004; Sugimoto 2000). For example, IRTIs were recorded to measure skin temperature variation around certain parts of the face for examining the autonomic nervous activity (Matsuzaki and Mizote 1996). The study suggested that fluctuations in facial temperature could provide a noninvasive measure to examine the autonomic nervous activity. Thermal facial screening was employed to detect attempted deceit using a three-stage system (Pavlidis, 2004). In the first stage of the system, thermal images were acquired using mid-range thermal equipment. Acquired images were used to transform facial thermal data into a blood flow model in the second stage. Such hemodynamic model was built upon the premise that significant blood flow redistribution takes place with a change in emotional condition and level of anxiety. During the third stage, the hemodynamic model was used to classify people into deceptive or nondeceptive categories. The system reportedly achieved results compatible with the polygraph examination by human experts (Khan et al., 2005). Figure 5.4 shows human face FTFPs, mapped to facial muscle FTFPs, obtaining the geometric profile of the FTFPs to the facial muscle shown in Table 5.1.

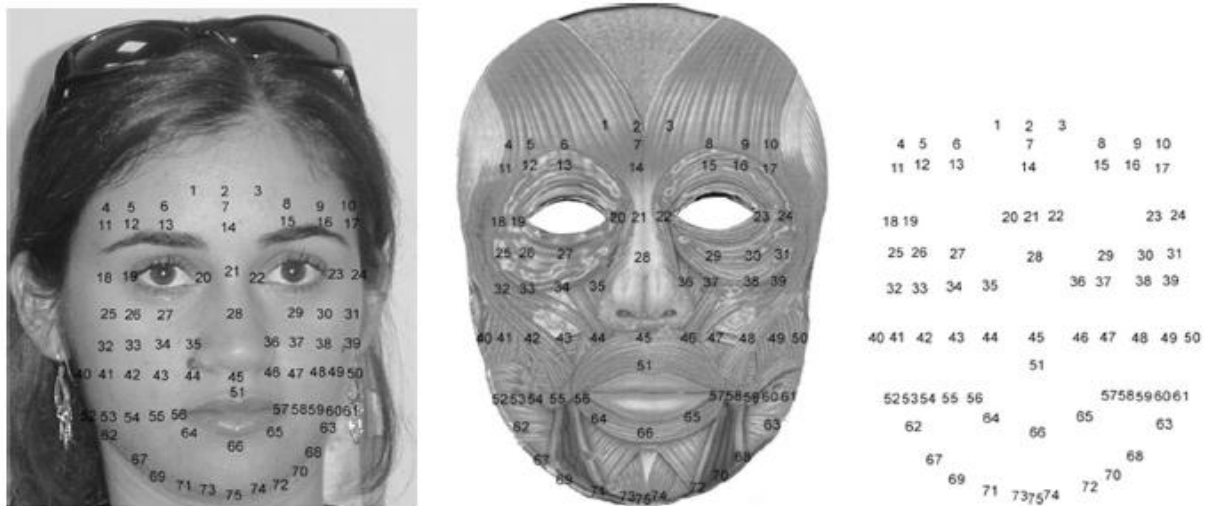


Figure 5.4 FTFPs on human face, facial muscle map, and a geometric profile of the FTFPs. (Source: KHAN, 2006).

Table 5.1 Muscular alignment of FTFPs. (Source: KHAN, 2006).

| Facial Muscle | Facial Thermal Feature Points (FTFP) |
|---|--|
| Frontalis, pars medialis | 1, 3, 6, 8, 13, 15 |
| Frontalis, inner center edges of pars medialis and pars lateralis | 2, 7 |
| Frontalis, pars lateralis | 4, 5, 9, 10, 11, 12, 16, 17 |
| Procerus/Levator, labii superioris alaeque nasi | 21 |
| Depressor, supercillii | 14 |
| Orbicularis Oculi, pars orbital | 18, 19, 20, 22, 23, 24, 25, 26, 27, 29, 30, 31 |
| Orbicularis Oris | 45, 51, 64, 65, 66 |
| Levator, labii superioris alaeque nasi | 28, 35, 36 |

5.2. Implementation

In this study of thermal facial variation, a four-stages system is developed. In the first stage of the system, thermal images are acquired using a Therm-App thermal equipment, whose acquired images are used to segment the facial thermal data into RoIs in the second stage. During the third stage, thermal features are extracted, and in the fourth stage, thermal feature variations are analyzed to estimate arousal and valence changes. Figure 5.5 shows a block diagram of the system implemented.

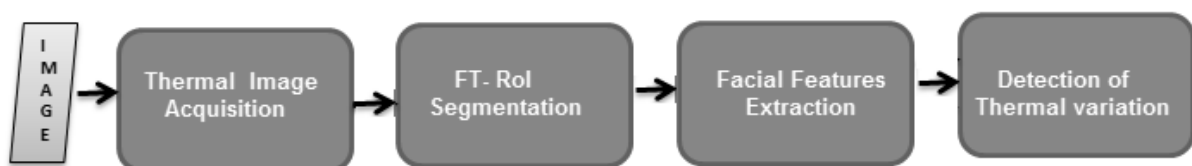


Figure 5.5 block diagram of the system here developed for thermal facial variation detection.

5.2.1. Data acquisition and segmentation

The data acquisition module allows acquiring thermal images from the Therm-app camera. The acquisition use the Therm-App software which is configured in night vision; the thermal image can be saved automated with a application developed in processing using REDIS DB; also the thermal image can be saved manually from a tablet. The whole process is recorded with the thermal camera, and the videos are saved for further analysis. Figure 5.6 shows an example of the acquisition process.

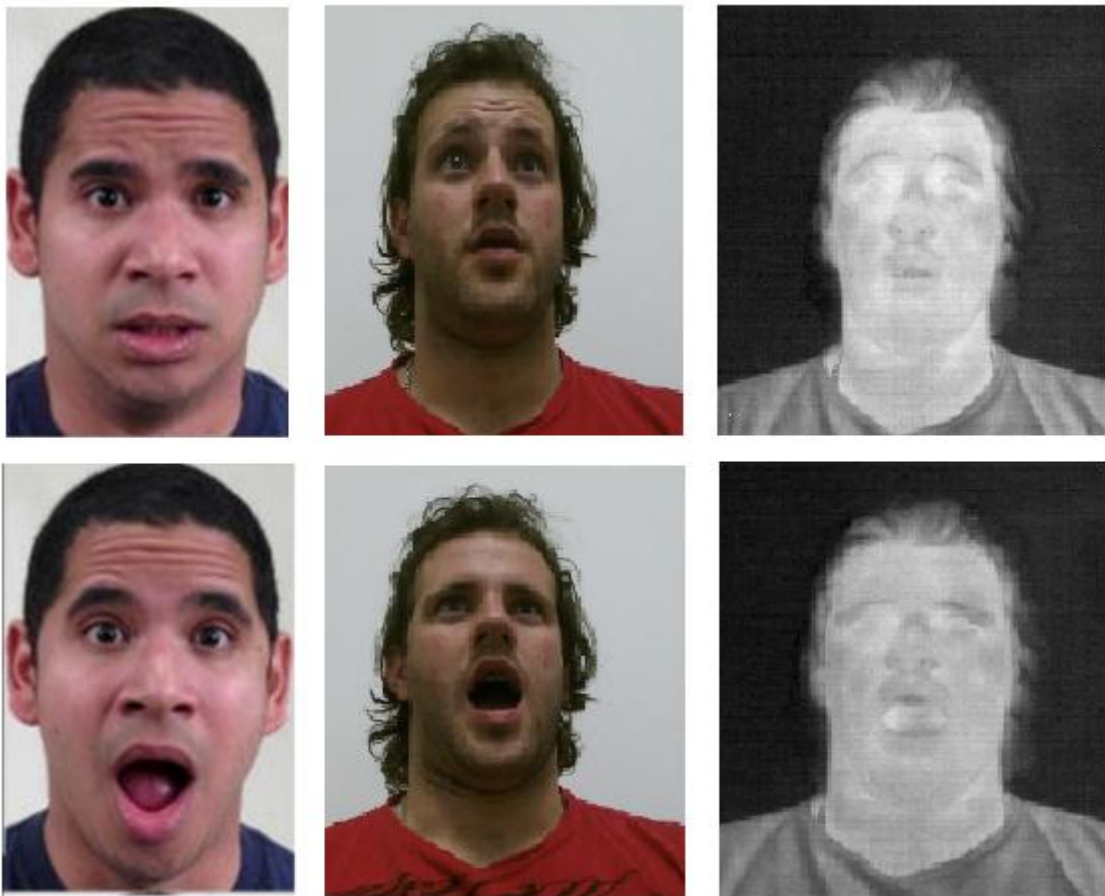


Figure 5.6 Example of the thermal image acquisition process.

The segmentation module allows obtaining the six FT-RoI (nose, chin, right cheek, left cheek, right forehead and left forehead), such as shown in Figure 5.7. To segment the image, it is used thresholds and morphological closing and opening filters. Then, in the segmented region, the six RoIs are placed. It is possible to obtain RoIs manually, in which the user selects the RoIs location (it is more accurate, but needs more time) or in automatic mode, where the system geometrically places the RoIs (faster, but not so precise, especially when the volunteer moves the head and rotates it in one of the three spacial axes).

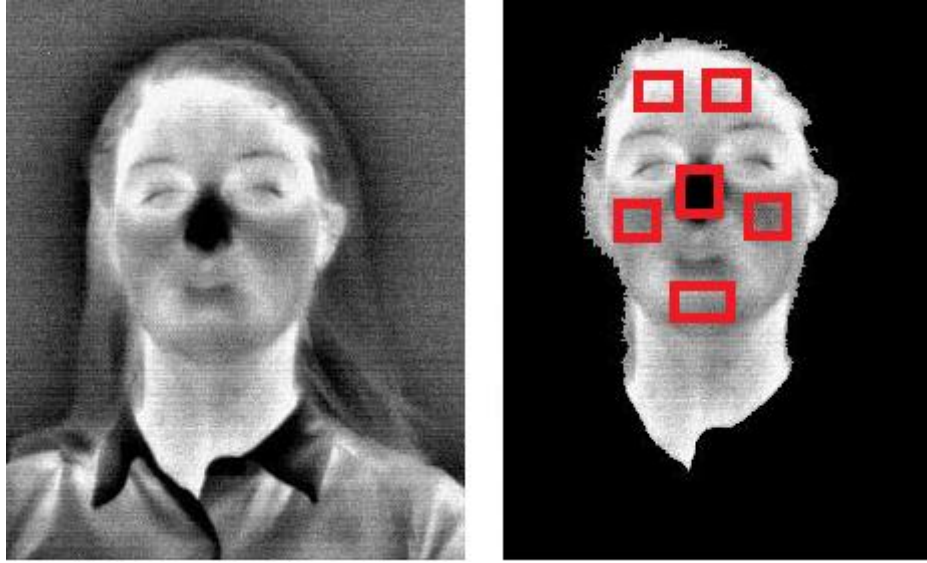


Figure 5.7 Example of the RoIs segmentation.

5.2.2. Features extraction and detection of thermal variation

The features extraction module allows obtaining thermal features from the RoIs. For each RoI, we obtain the median of five temperature measurements from different frames to filter the noise of the image, and from these measures the features are calculated. For each RoI, the difference to the corresponding baseline is calculated ($\text{RoI} - \text{BaseLine}$), obtaining seven features, one for each RoI, plus the facial average temperature and the difference of each RoI regarding the facial average temperature ($\text{RoI} - \text{Facial Temperature}$), which implies obtaining six more features, for a total of thirteen features. Figure 5.8 shows the graphics of the thirteen features used in this work.

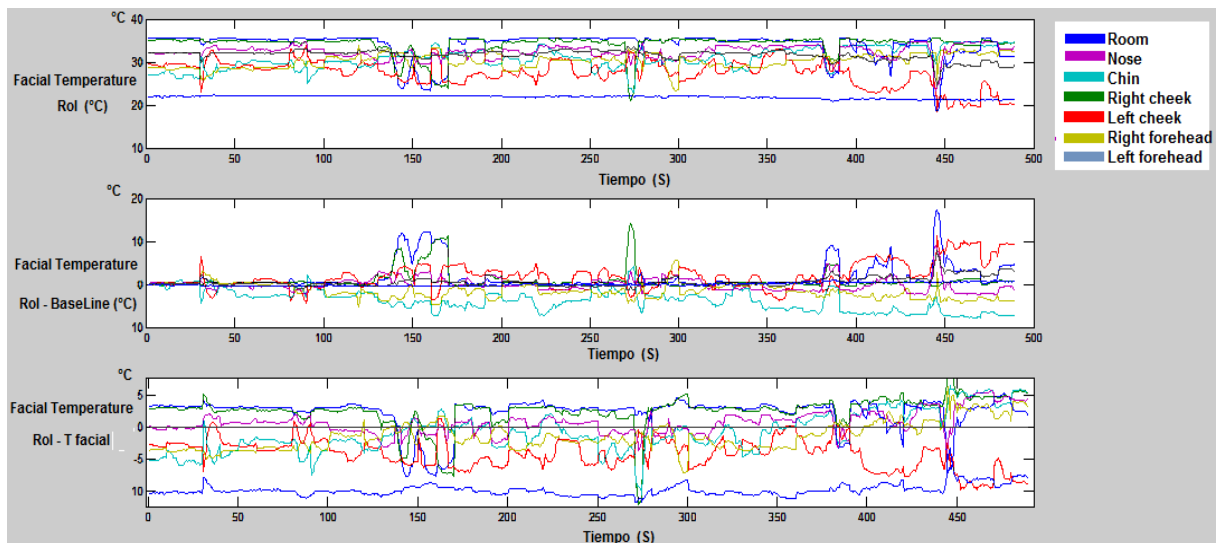


Figure 5.8 Features used in this work (RoIs temperature, $\text{RoI} - \text{BaseLine}$, and $\text{RoI} - \text{Facial Temperature}$).

With the thermal variation module, it is possible to process images of the videos obtained from the experiments. The images can be processed in order to get the volunteer's thermal features at each experiment, making possible to analyze temperature variation patterns during evoked emotions. Figure 5.9 shows the analysis of one image and the corresponding pattern of thermal features (bar 1 to 7 RoI- BaseLine and bar 8 to 13 RoI – Temperature facial).

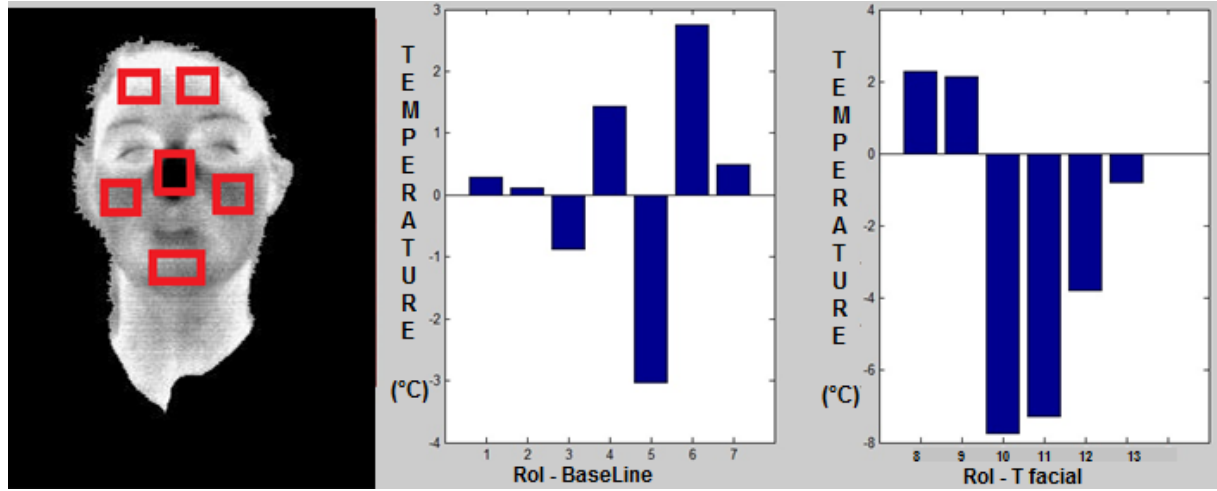


Figure 5.9 Features extraction (RoI - BaseLine and RoI - Facial Temperature).

5.3. Analysis and results

The Therm-App in night mode does not provide a linear measure of temperature (Therm-App, 2014), so results and tables are presented in percentage variation related to the maximum value. The Equation (5.1) shows the percentage variation.

$$\%Variation = \frac{Variation \times 100}{\max Value} \quad (5.1)$$

Procedure 1: Variation of facial temperature by facial expressions

In this experiment, each volunteer is asked to sit comfortably in a chair positioned in front of a 19-inches computer screen and a Therm-App sensor, with his/her eyes at 80 cm away from the screen, and at 70 cm from the sensor. The screen displays the six pictures related to human facial expressions for ten seconds. The volunteer should imitate each emotional expression three times. The Therm-App sensor records images of each emotional facial expression performed by the volunteer, and an algorithm identifies the set of features related to expressions of each emotion, based on thermal variation for the six basic facial expressions. Figure 5.10 shows thermal images for the six facial expressions (fear, sadness, anger, happiness, surprise and disgust).

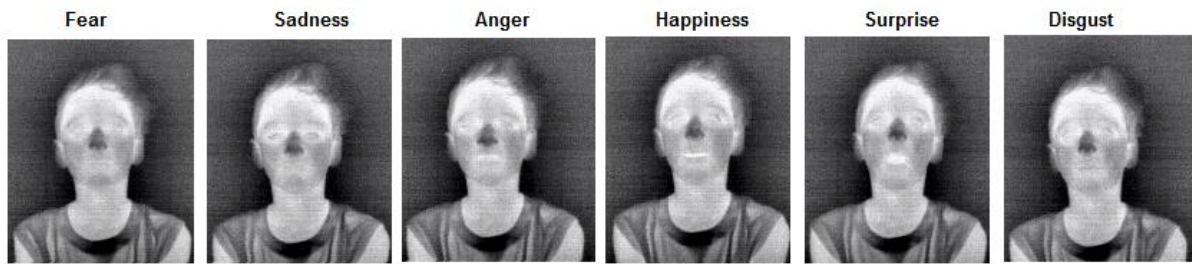


Figure 5.10 Thermal images for the six facial expressions considered in this work.

Table 5.2 shows the results obtained for the thermal variation of RoIs for the different facial expressions. The thermal variation in the regions is very small, less than 2%; only the nose has variation from 3% to 5%, but it is not possible to recognize any pattern in the results. It is worth to mention that the sensor sensitivity is 0.07°C , while the temperature variation for facial expressions is about 0.05°C

Table 5.2 Percentage of RoIs thermal variation for facial expressions

| Expression | Nose | Chin | Right cheek | Left cheek | Right forehead | Left forehead |
|------------|------|------|-------------|------------|----------------|---------------|
| Anger | 3% | 1% | 1% | 1% | 1% | 1% |
| Fear | 3% | 1% | 1% | 1% | 1% | 1% |
| Sadness | 5% | 2% | 3% | 3% | 1% | 1% |
| Happiness | 3% | 1% | 1% | 1% | 2% | 2% |
| Surprise | 3% | 1% | 2% | 2% | 1% | 1% |
| Disgust | 1% | 1% | 1% | 1% | 1% | 1% |

Procedure 2: Variation of facial temperature by emotions (arousal and valence)

This procedure consist of the visualization by the volunteer of the six emotion-inducing videos to evoke certain emotions (surprise, sadness, disgust, fear and happiness). Each volunteer sit down comfortably in a chair in front of both a screen and a box with the camera system. The Therm-App sensor records images of each emotional facial expression performed by the volunteer, and an algorithm identifies the set of features related to expressions of each emotion, based on thermal variation for the valence and arousal detection. Figure 5.10 shows thermal images for different valence. In the experiments, a great temperature variation at the nose region was observed for variation in positive and negative valence stimuli, while the other regions of the face did not show a large variation to these stimuli.

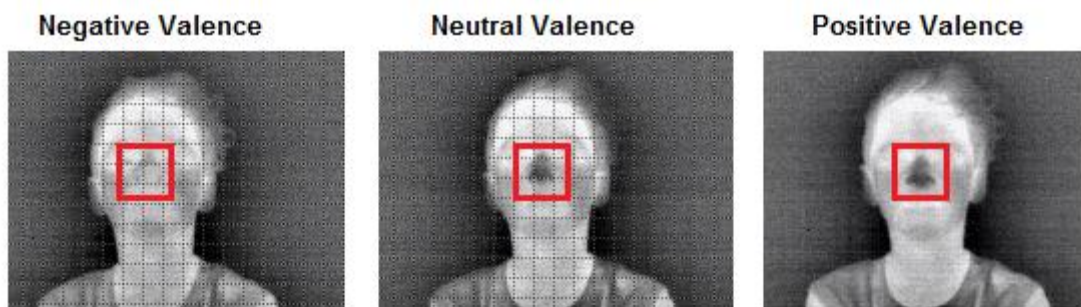


Figure 5.10 Thermal images for negative, neutral and positive valence

Figure 5.11 shows thermal images for different arousal stimuli. In the experiments, a variation in the region of the forehead and cheek was observed for variation in the arousal stimuli, as the temperature in these regions increases when the arousal increases.

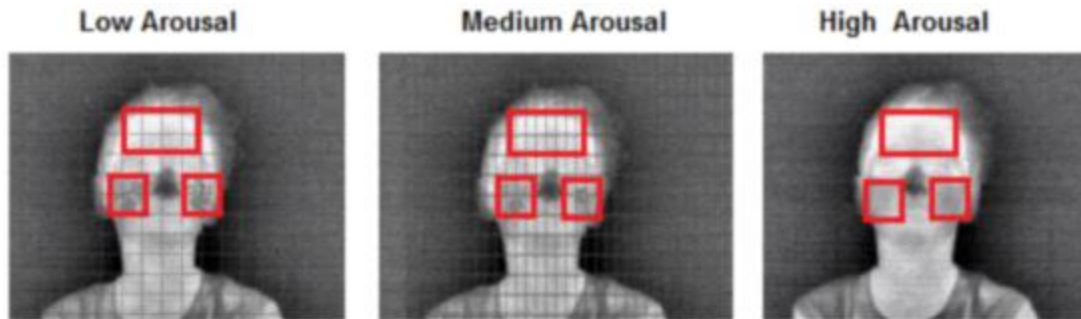


Figure 5.11 Thermal images for low, medium and high arousal.

Table 5.3 shows the results from the experiments. For valence, the most important variation was observed in the nose; for positive valence, the temperature decreases up to 4%, and for negative valence, it can increase up to 8% respect to images of neutral valence. On the other hand, for arousal, an increase from 3% to 4% in cheek and forehead temperature was observed; the nose also had a slightly lower increase of 2%.

Table 5.3 Percentage of RoIs thermal variation in arousal and valence.

| Emotion | Nose | Chin | Right cheek | Left cheek | Right forehead | Left forehead |
|------------------|------------|------|-------------|------------|----------------|---------------|
| Positive Valence | -4% | 2% | 2% | 2% | 3% | 3% |
| Neutral Valence | 1% | 0% | 0% | 1% | 1% | 0% |
| Negative Valence | +8% | 2% | 2% | 2% | 3% | 3% |
| Low Arousal | 1% | 0% | -1% | -1% | 0% | 0% |
| High arousal | +2% | 1% | +3% | +3% | +4% | +4% |

5.4. Discussion

In this chapter, a system was developed in order to analyze facial thermal variation based on the technique of Facial Thermal-Region of Interest FT-RoI. The developed system allows: data and thermal image acquisition, Region of Interest (FT-RoI) segmentation; features extraction; and, detection of facial thermal variation. For the different facial expressions, the thermal variation in the regions is very small, and it is not possible to recognize facial expressions based on the measured thermal variation. However, for valence and arousal, it was found a relation between facial thermal variation and emotional activity. This system and its implemented functions are the basis for the applications of emotion recognition, which are presented in Chapters 6 and 7.

The data acquisition module here developed is able to acquire thermal images from the Therm-app camera, which has a resolution of 384 x 288 pixels, accuracy of $\pm 3^{\circ}\text{C}$, sensitivity of 0.07°C and temperature range of 5 to 90°C , and capture mode in night vision (Therm-App, 2014). Therefore, the results obtained in our research are limited by these features of the sensor. The segmentation module gets six FT-RoI (nose, chin, right cheek, left cheek, right forehead and left forehead). Based on the FT-RoIs, the feature extraction module obtains thirteen thermal features (seven features for RoI - BaseLine and six more features for RoI - Facial Temperature). Finally, the thermal variation module obtains the thermal features of the volunteer, in order to study possible patterns in the variation of temperature in situation of emotional variation.

The results obtained for the first proposed experimental procedure are that it is not possible to obtain a sufficiently small measure of temperature, as the sensitivity of the sensor is not small enough for this analysis. Thus, with this sensor, it is not possible to analyze facial thermal changes due to facial expressions since the thermal variation of facial expressions is about 0.05°C , while the minimum sensitivity of the sensor is 0.07°C .

The results obtained for the second proposed experimental procedure are that the temperature patterns of the RoIs show variation for different valence and arousal stimuli. For valence, there are variations in the nose, and for arousal there are changes in the forehead and cheeks. We consider that these results are not conclusive, maybe due to the few number of participants or because the stimuli were not sufficiently strong neither long as to produce strong emotion variation.

CHAPTER 6

6. MULTISENSORIAL INTEGRATION

One of the challenging issues in affective computing is to endow a machine with an emotional intelligence. Humans employ multiple sensors in emotion recognition. At the same way, an emotionally intelligent system requires multiples sensors to be able to create an affective interaction with users. Many factors render multisensorial emotion recognition approaches appealing. First, humans employ a multisensorial approach in emotion recognition, then, machines attempt to reproduce elements of the human emotional intelligence. Second, the combination of multiple-affective signals not only provides a richer collection of data, but also helps alleviating the effects of uncertainty in the raw signals.

In this Chapter, three multisensorial integration strategies are proposed and implemented: Kinect and eye tracker integration, thermal camera and Kinect integration, and thermal camera, Kinect and eye tracker integration. Figure 6.1 shows a multisensorial integration of thermal camera, Kinect and eye tracker, in order to improve the results of the system.

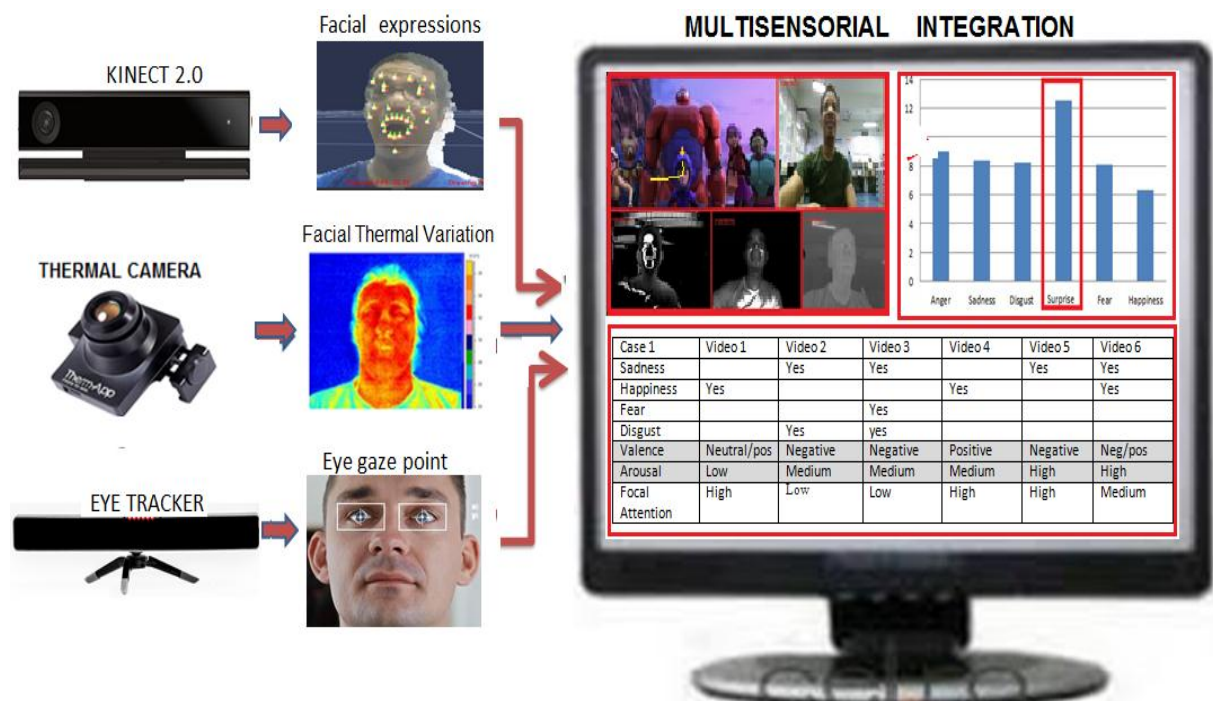


Figure 6.1 Multisensorial integration: thermal camera-Kinect-eye tracker.

6.1. Background Multisensorial Integration

The multisensorial approach here developed presents challenges associated with the fusion of single signals, dimensionality of the feature space, and incompatibility of collected signals in terms of time, resolution and format. With these multimodal emotion recognition approaches, information extracted from each modality are reconciled to obtain a single emotion classification result, which is known as multimodal integration. The literature on this topic is rich and generally describes three types of integration mechanisms: feature-level integration, decision-level integration, and hybrid approaches.

6.1.1. Feature-level Integration

A common method to perform modality integration is to create a single set from all collected features, and a single classifier is then trained on the feature set. However, feature-level integration is plagued by several challenges: first, multimodal feature set contains more information than a unimodal one, which can present difficulties if the training dataset is limited. In fact, Hughes (1968) has proven that the increase in the feature set may decrease classification accuracy if the training set is not large enough. Second, features from various modalities are collected at different time scales (Pantic, 2003). For example, features of Heart Rate Variability HRV in frequency domain typically summarizes seconds or minutes' worth of data (Al Osman, 2016), while speech features can be in the order of milliseconds. Third, a large feature set undoubtedly increases the computational load of the classification algorithm (Lingenfelter, 2011). Finally, one of the advantages of multimodal emotion recognition is the ability of synchronizing data easily and producing an emotion classification result in the presence of missing or corrupted data. However, feature-level integration is more vulnerable to the latter issues than decision-level integration techniques (Wagner, 2011).

6.1.2. Decision-level Integration

Typically, an emotion recognition system produces errors in some area of the feature space (Alexandre, 2001). Hence, combining the results of multiple systems can alleviate this shortcoming. This is especially true when each system is operating on a different modality that corresponds to a separate feature space. Using decision-level integration, modalities can be independently classified using separate models, and the results are joined using a multitude of possible methods. Therefore, this approach is said to employ an ensemble of systems and classifiers. Ensemble members can belong to the same family or different families of statistical classifiers. In fact, static and dynamic classifiers can both be employed in such a multimodal system.

6.1.3. Hybrid-level integration

When an integration technique combines feature and decision-level integration, it is referred to as a hybrid- integration scheme. For instance, we can achieve integration in two stages. In the first stage, a system can perform feature-level integration. For example, a single classifier can handle features from audio and video signals. In the second stage, decision-level integration can be used to combine the results of that with another one operating on physiological (e.g., HRV) features. Kim (2005) proposes a simple hybrid- integration approach where the result from the feature-level integration is fed as an additional input to the decision-level integration stage.

6.2. Implementation of a multisensorial system for emotion recognition.

Figure 6.2 shows the block diagram of the three integration levels implemented in this research. In the first stage of the system, eye tracker and Kinect were integrated using a decision-level technique, and a feature-level technique was used to integrate thermal camera and Kinect in the second stage. During the third stage, a hybrid-level technique was used to integrate thermal camera, eye tracker and Kinect.

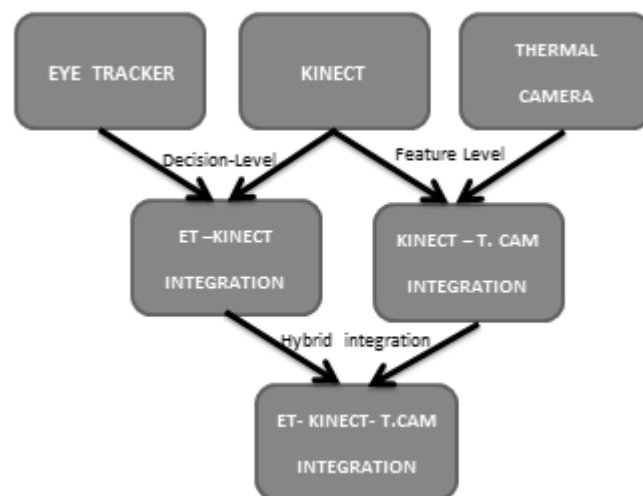


Figure 6.2 Block diagram of the proposed integration strategy

6.2.1. Data-Level Integration on Processing Language

The first type of integration attempted in this research was at the data level. A system was implemented in Java-Processing, which acquired the data of the three sensors in a single application. The advantage of this proposal was that the data were acquired synchronously and processed online. But the disadvantage was the processing consumption required by each sensor. Kinect requires USB3.0 technology to acquire data from its three cameras (color, infrared and depth). Eye tracker also uses USB3.0 and a complex processing algorithm, while

the thermal camera requires virtualizing the android operating system and a Redis server to transfer images.

This integration is done in a single computer (NUC). As the requirements of the sensors exceeded the characteristics of the previous computer, therefore, when one of the sensors failed the whole system failed. Due to continuous failures in this centralized technique, then one opted to abandon it and propose other decentralized technique. Figure 6.3 shows images acquired and processed online in the centralized application developed in processing.



Figure 6.3 Data-level integration online for data processing.

6.2.2. Decision-Level Integration eye tracker and Kinect

The integration of eye tracker and Kinect allows to carry out studies about focus of attention, in order to evaluate which parts of the face people focus on when they come to recognize expressions, and what is the stimulus that generates an emotional reaction in a person. In Chapter 7 we show the use of this integration.

Integration at the decision level was used in our research, since in this integration the two systems do not share features to obtain results, only the final result of the focus of attention obtained by the eye tracker with the result of expression recognition obtained by the Kinect are integrated. Figure 6.4 shows the integration where the focus of attention is detected on a face of a person to recognize a facial expression.

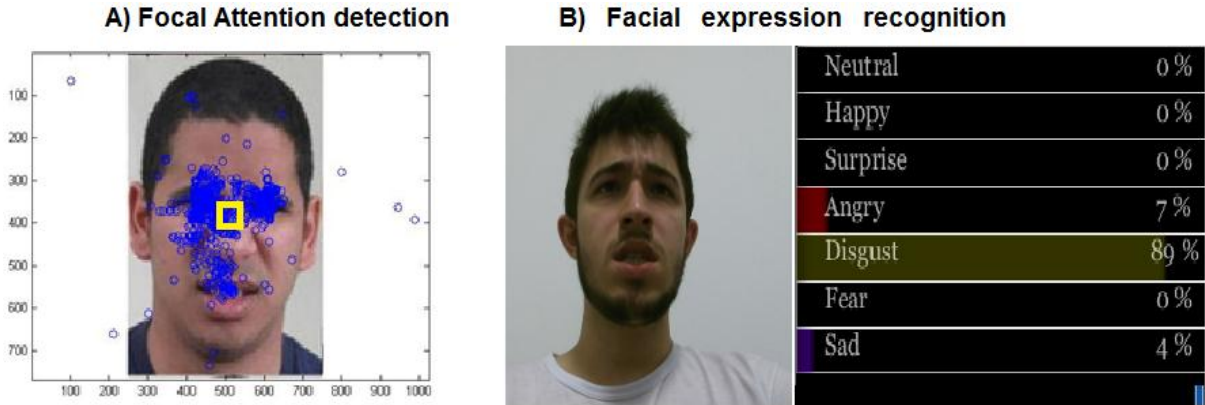


Figure 6.4 Eye tracker- Kinect integration: A) Focus of attention detection, B) Facial expression recognition

6.2.3. Feature-Level integration Kinect and thermal camera

The integration of Kinect and thermal camera allows improving the thermal feature detection. The main problem to detect these thermal features of the face is the difficulty of segmenting the RoI, because in a thermal image it is difficult to detect regions such as eyes, mouth or nose. In contrast, in the color image, these facial features are very easy to detect. The approach used in our research was to take the features obtained from the Kinect (AUs), and to project them in the thermal image. Projecting the points of the AU in the thermal image is easier, in order to automatically segment the RoI and obtain the thermal characteristics.

A mathematical model for these (AUs) projection was implemented using a camera calibration toolbox in Matlab. Any three-dimensional point (X_W, Y_W, Z_W) in the scene can be taken to a camera coordinate system (X_C, Y_C, Z_C), which is achieved with a rotation matrix R and translation vector T , such as shown in Equation (6.1):

$$\begin{pmatrix} X_C \\ Y_C \\ Z_C \end{pmatrix} = R \begin{pmatrix} X_W \\ Y_W \\ Z_W \end{pmatrix} + T = \begin{pmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{pmatrix} \begin{pmatrix} X_W \\ Y_W \\ Z_W \end{pmatrix} + \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix} \quad (6.1)$$

The values of R and T are known as extrinsic parameters, then this coordinated system in space must be taken into the two-dimensional space of the images, without taking into account the radial and tangential information. The coordinates of the image (x_b, y_b) are shown in Equation (6.2).

$$\begin{pmatrix} x_b \\ y_b \\ 1 \end{pmatrix} = k \begin{pmatrix} f_{cx} & s & C_x \\ 0 & f_{cy} & C_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_e \\ Y_e \\ Z_e \end{pmatrix}$$

(6.2)

where:

f_{cx} and f_{cy} are the focal distances expressed in pixels, and include the focal length of the camera and the size in millimeters of the sensor (S_x , S_y); and C_x and C_y are the optical center of the image. The value s is called framing, and most of the time it corresponds to an angle of 90° and therefore its value is zero. The value k is a scaling factor, and the values f_{cx} , f_{cy} , C_x , C_y , s , k are known as intrinsic parameters. Figure 6.5 shows the calibration process for AUs projection on the thermal image. The AUs points of the Kinect image are projected to its corresponding 3D model, then this model is transformed to the 2D model on the thermal image (Figure 6.6).

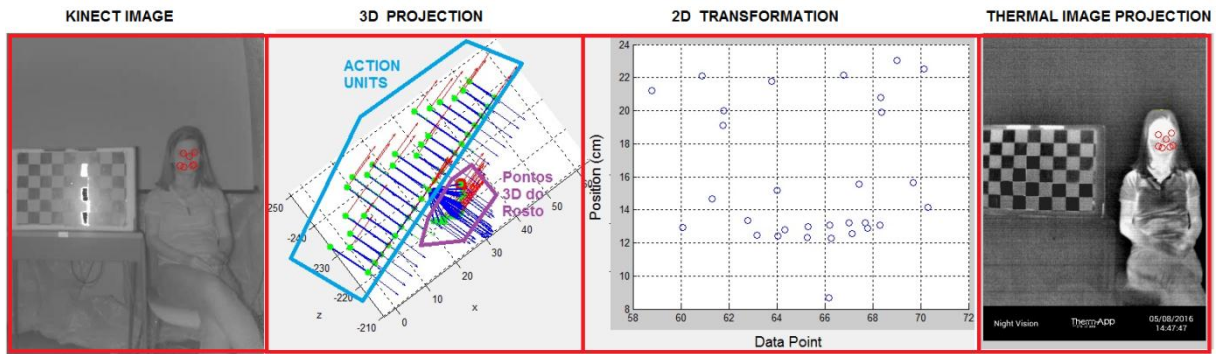


Figure 6.5 Calibration process for AUs projection on the thermal image.

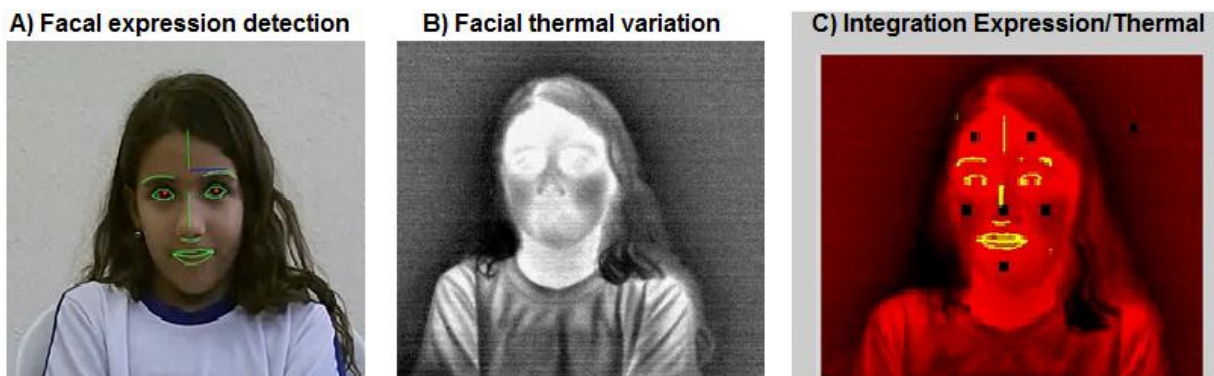


Figure 6.6 Projection of AU points from Color to Thermal image: A) Facial expression detection; B) Facial thermal variation; C) Integration of AUs on thermal image.

The integration of Kinect and thermal camera improves the thermal feature detection and the RoI segmentation. In Chapter 7 the use of this integration is shown.

6.2.4. Hybrid-Level Integration: eye tracker, Kinect and thermal camera

For the integration of the three systems, a hybrid-level integration was proposed to take advantage of the two integrations previously done (Kinect-eye tracker and Kinect-camera). The purpose of the integration of the three systems is to be able to give a more complete evaluation of the emotional state during the experimental stage by integrating the results of focus of attention, recognition of facial expressions and emotional variation of arousal and valence. The multisensorial integration allows realizing studies of social focal attention, recognition and expression of emotions, and to detect variation of the emotional state of a person. In Chapter 7 the experimental part is shown and the results are explained. Figure 6.7 shows the multisensorial integration implemented in this work, allowing focal attention detection, facial expression recognition, and estimation of emotional state.

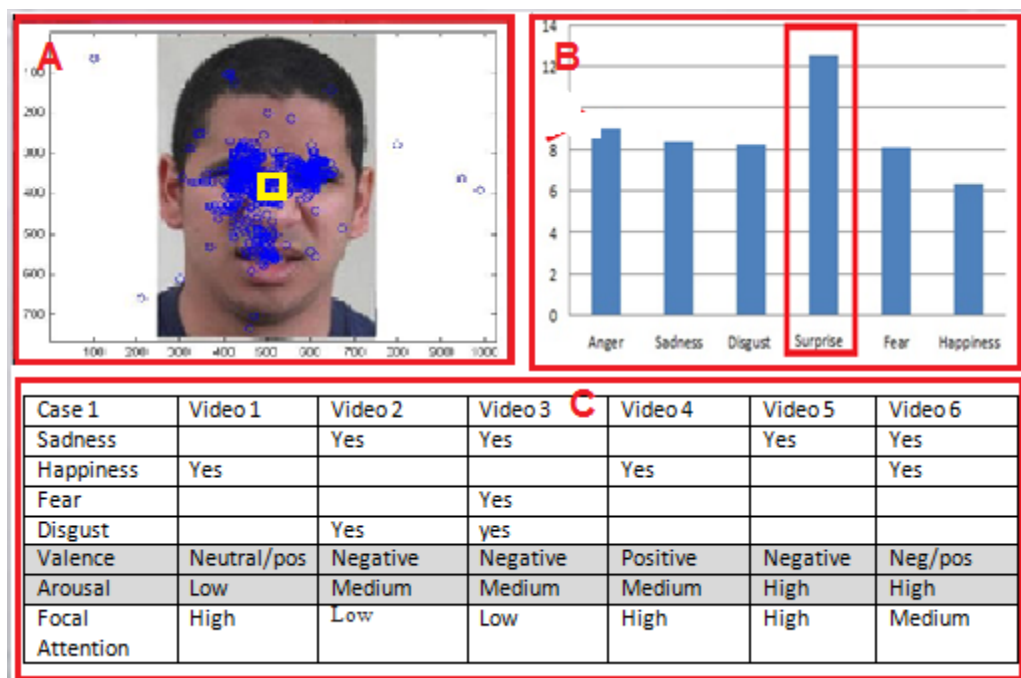


Figure 6.7 Multisensorial integration: A) Focal attention detection, B) Facial expression recognition, C) Estimation of emotional state.

The experimental results and the discussion of the developed multisensory system are presented in Chapter 7.

CHAPTER 7

7. VALIDATION OF MULTISENSORIAL SYSTEM

Multisensorial emotion recognition methods require multisensorial systems to collect the relevant data from expressions, as these systems are more complex than the unisensorial ones in terms of the number and diversity of sensors involved, and computational complexity of the data-interpreting algorithms. This challenge is more evident when data are analyzed, since it is necessary to synchronize the data of each sensor and show integrated results that allow a better analysis than the unisensorial results.

In this chapter, the results from multisensorial emotion recognition are presented. Three experimental procedures were developed using the platform and environment for the experiments presented in Section 2.3. The first experimental procedure was designed to evaluate social visual attention; the second procedure was proposed to evaluate the recognition of facial expressions and emotional variation; and the third procedure was designed to evaluate emotions by integrating the three sensors (eye tracker, Kinect and thermal camera). Finally, the results are compared with the functional and technical requirements of the research presented in Section 1.1.2. Figure 7.1 shows the three types of stimuli used in the experiments.

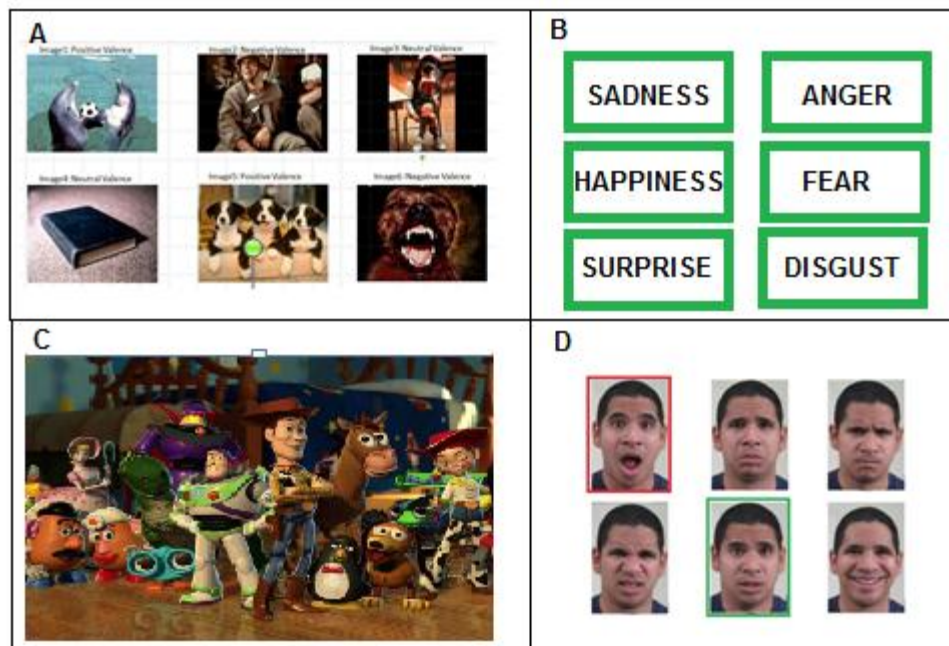


Figure 7.1 Stimuli used in the three experiments conducted in this research; A) images for valence study; B) names of the basic emotions; C) emotion-inducing videos; D) pictures relative to human facial expressions.

7.1. Experiment 1: Social Focal Attention Recognition

This research approaches a procedure used to assess visual attention through the displaying of pictures of positive, negative and neutral valence.

7.1.1. Experimental Protocol

This procedure has the participation of sixteen healthy adult volunteers (twelve men and four women), with mean age of 28 years old (± 5.32). Each volunteer is invited to sit comfortably in a chair positioned in front of the screen of a computer (19 inches) and an eye tracking device (Eye Tribe), with eyes at 70 cm from screen and at 60 cm from eye-tracker. The volunteer hears a brief explanation about the procedure and solves him/her doubts. Figure 7.2A shows the setup used for the experimental test. A previous calibration is necessary to gather a good data acquisition, which consists of tracking visually mobile points in the screen and, subsequently, fixating points of known coordinates in the viewing scene.

The participant views a set of six images (1024 x 768 pixels), being two of positive valence, two of negative valence and two neutral. The valence classification is based on 1-9 scale, where scores were > 7 (for pleasant pictures), < 5 (for unpleasant pictures) and between 5 and 6.5 (for neutral pictures), respectively. Figure 7.2B shows the chosen images portray, with puppies and animals playing, for positive valence (images 1 and 5); injured person and angry animal, for negative valence (images 2 and 6), and person in daily activities and common objects, for neutral valence (images 3 and 4). The pictures are selected from an international database (IAPS – International Affective Picture System), commonly used in studies about emotions and attention (Lang 2008). The picture set is displayed five times and the time of exhibition of each image is ten seconds.

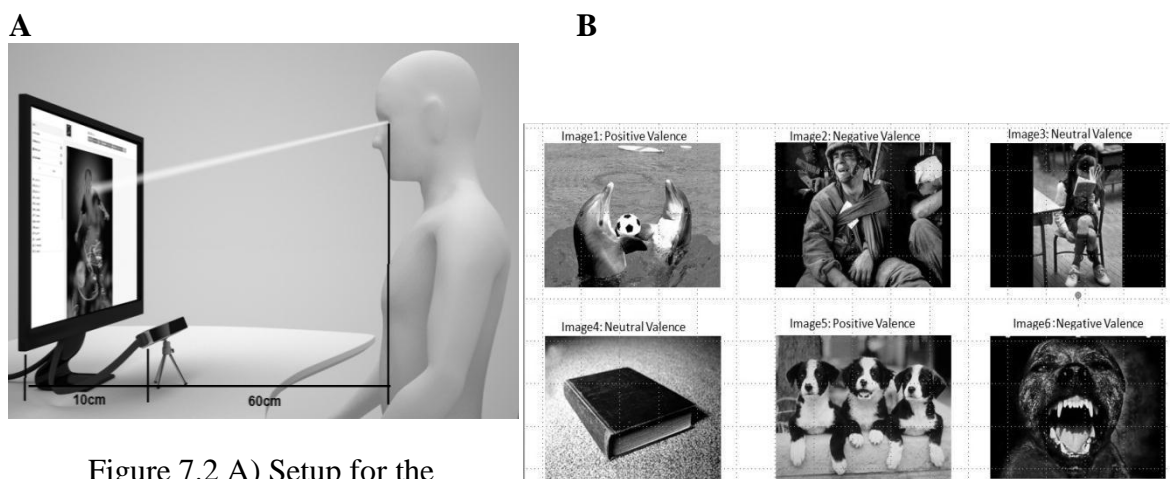


Figure 7.2 A) Setup for the experimental tests. B) Set de images for Valence Study. Source: IAPS (Lang 2008)

In this procedure, it is assessed the number of times and the time of viewing of the pictures, in order to identify which valence stimulus (negative, positive or neutral) got more attention.

7.1.2. Results

Table 7.1 indicates the percentage of the average time of viewing each image by the participants. These images correspond to valence: positive (images 1 and 5), negative (images 2 and 6) and neutral (3 and 4). The highest values and the smallest values are bolded. Table 7.2 shows the number of observers who presented highest and lowest attention time to positive and negative valence and neutral stimuli.

Table 7.1: Percentage of the time of viewing of the pictures.

| Picture Number | Time of viewing (%) |
|-------------------|---------------------|
| Image1 | 19.96 |
| Image2 | 16.33 |
| Image3 | 15.52 |
| Image4 | 11.49 |
| Image5 | 19.33 |
| Image6 | 10.30 |
| Outside of images | 7.04 |

Table 7.2. Number of observers who present highest and lowest attention to pictures featured by the valence.

| Valence | Maximum attention attracted (Number of people) | Minimum attention attracted (Number of people) |
|----------|--|--|
| Positive | 11 | 1 |
| Negative | 3 | 7 |
| Neutral | 2 | 8 |

Images 1 e 5, which correspond to the positive valence, have the highest percentage of average time of viewing, with 19.96% and 19.33%, respectively. On the other hand, image 6, which corresponds to the negative valence, has the lowest percentage of average time of viewing with 10.30%. From Table 7.2, the images with positive valence had high number of observers (11 participants), whereas images with negative valence had 3, and neutral images had 2. The images with neutral and negative valence elicited low attention in 8 and 7 participants, whereas images with positive valence elicited low attention in 1 participant.

7.2. Experiment 2: Expression comprehension and recognition

7.2.1. Experimental Protocol

This procedure has the participation of eleven healthy adult volunteers (eight men and three women), with mean age of 28.27 years old (± 5.33). The participant's setup in this test are the same as described in Procedure 1. In the first exhibition, the participant views six pictures relative to human facial expressions (surprise, sadness, anger, disgust, fear and happiness) for 10 s, individually. Then, the participant should answer the emotion correspondent to the viewed emotional expression. Figure 7.3 shows examples of human face emotional expressions used in the procedure.

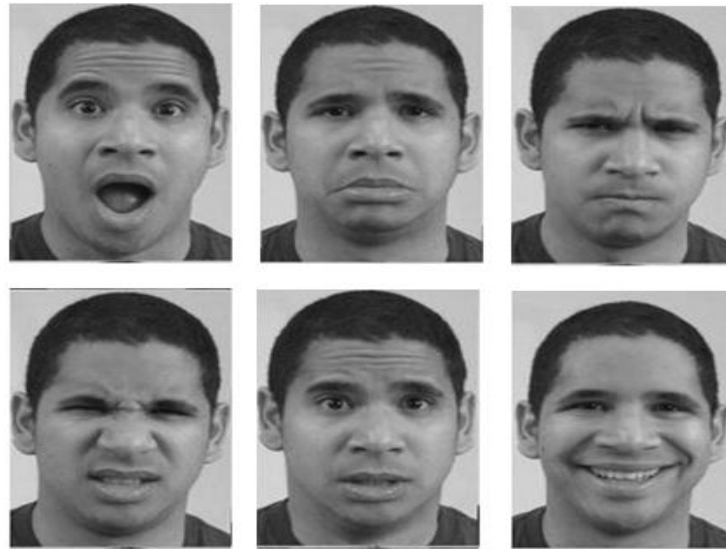


Figure 7.3 Examples of human face emotional expressions used in the procedure 2. (Source: Du, 2014)

In sequence, the volunteer views the set of the six human affective pictures for 3 times, during 10 s, displayed together, as observed in Figure 7.3. Finally, the volunteer is asked to focus on the picture (among the six) correspondent to the emotion said by the mediator of the procedure. Then, when the participant focus rightly, the border of the picture become green. In case of wrong focus, the picture border become red.

With this procedure, it is evaluated: a) which face regions the participant focused on to recognize an emotion and if he/she identifies the emotion correctly; b) which emotional facial expressions more attracts his/her attention; and c) if the participant has difficulty to recognize the emotion required by the mediator. For this, is assessed the number of times and the time of viewing the pictures.

7.2.2. Results

Figure 7.4 shows an example of data obtained from the eye-tracking sensor (blue circles). These data were processed to detect the regions of the pictures more observed (attention focus), during the recognition of emotions in the facial expressions. The red square in Figure characterizes the mean attention focus.

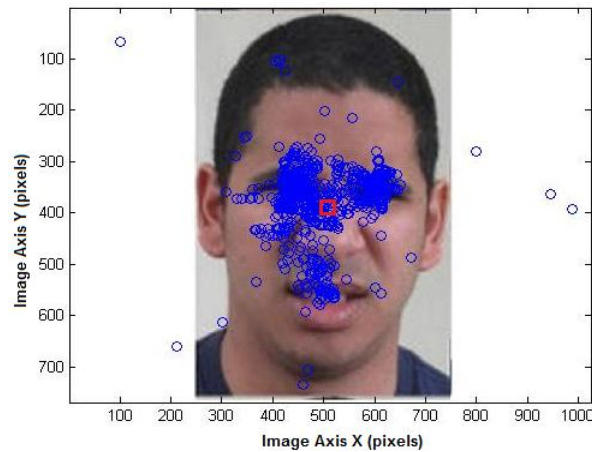


Figure 7.4: Data from eye-tracking sensor referent to attention focus, featured by blue circles overlapping on the facial image. The mean focus obtained is shown in red square.

Table 7.3 shows the result of the calculation of the average and standard deviation in pixels for the focus of attention performed for all participants in relation to all pictures relative to emotional facial expressions.

Table 7.3 Mean and standard deviation of the focus points performed by the participants during the visualization of facial expressions.

| Screen axis | Center of the Screen (pixels) | Attention focus (pixels) | Standard deviation (pixels) |
|-------------|-------------------------------|--------------------------|-----------------------------|
| Axis X | 512 | 498.8 | 20.4 |
| Axis Y | 384 | 385.8 | 19.1 |

Table 7.4 shows the average time required by the participants to recognize the six facial expressions. The highest and the lowest values are bolded.

Table 7.4 Time to recognize the emotional facial expressions.

| Emotion expressions | Time to recognize (s) | Standard deviation (s) |
|---------------------|-----------------------|------------------------|
| Anger | 9.71 | 5.08 |
| Sadness | 8.72 | 2.49 |
| Surprise | 8.79 | 3.88 |
| Disgust | 12.50 | 7.24 |
| Fear | 8.89 | 3.52 |
| Happiness | 6.33 | 0.60 |

Table 7.5 shows the number of mistakes of the participants in the recognition of emotions from facial expressions.

Table 7.5 Number of mistakes in the facial expressions recognition.

| Emotion expressions | Number of mistakes |
|---------------------|--------------------|
| Anger | 1 |
| Sadness | 1 |
| Surprise | 2 |
| Disgust | 3 |
| Fear | 1 |
| Happiness | 0 |
| Total | 8 |

Table 7.6 indicates the percentage of expression recognition, the valence and arousal detected and the percentage of focus detection for the experiment.

Table 7.6 Values for expression recognition.

| Expression | Expression recognition | Valence detection | Arousal detection | Focus of attention |
|--------------|------------------------|-------------------|-------------------|--------------------|
| Anger | 58% | Neutral | Low | 83% |
| Sadness | 50% | Neutral | Low | 79% |
| Surprise | 85% | Neutral | Low | 72% |
| Disgust | 75% | Neutral | Low | 68% |
| Fear | 64% | Neutral | Low | 79% |
| Happiness | 89% | Neutral | Low | 87% |
| Total | 70% | Neutral | Low | 78% |

The highest focus of attention is relative to the central regions of the pictures (498.8x385.8 pixels) exhibited in the screen, featured by the regions of the eyes, nose and cheeks.

On the other hand, the average time required by the participants recognizing the six facial expressions was low for the happiness emotion (6.33 s) and high for the disgust (12.50 s). It was also noted that for the sixteen volunteers the number of mistakes for emotion recognition was higher for disgust (3 mistakes) while happiness had no mistake. Finally, the expression recognition was 70%, and for changes detected in the temperature, which would show variation in valence or arousal, the average of focus of attention was 78%.

7.3. Experiment 3: Multisensorial Emotion Analysis

7.3.1. Experimental Protocol

This procedure has the participation of 105 healthy children volunteers, with age ranged between 6 to 11 years old. The initial procedure consists of the visualization and the imitation of facial expressions from pictures relative to six classes of emotions. Each volunteer sits down comfortably in a chair in front of both a screen and a box with the camera system. The screen exhibits six names of the basic human emotions (surprise, sadness, anger, disgust, fear and happiness). Each emotion name is displayed during five seconds, allowing the participant making the emotional expression visualized. Afterwards, the screen exhibits six pictures of human faces expressing six emotional expressions (surprise, sadness, anger, disgust, fear and happiness). Each picture is displayed during five seconds, allowing the participant imitating the emotional expression visualized. Finally, the screen exhibits six emotion-inducing videos for evoke certain emotions (surprise, sadness, disgust, fear and happiness). Table 7.7 shows the emotions that each video intended to evoke. It is worth to comment that due to the difficulty in synchronizing data from the different sensors, three participants, from the database, were selected and evaluated as study case.

Table 7.7 Emotions that each video is intended to evoke.

| Case 1 | Video 1 | Video 2 | Video 3 | Video 4 | Video 5 | Video 6 |
|-----------------|----------|----------|----------|----------|----------|---------|
| Sadness | - | Yes | Yes | - | Yes | Yes |
| Happiness | Yes | - | - | Yes | - | Yes |
| Fear | - | Yes | Yes | - | - | - |
| Disgust | - | Yes | yes | - | - | - |
| Valence | positive | Negative | Negative | Positive | Negative | Neg/pos |
| Arousal | Low | High | High | High | High | Medium |
| Focal Attention | High | Low | Low | High | Medium | Medium |

7.3.2. Results

Case 1

Table 7.8 indicates the emotions recognized by the multisensorial system. For volunteer 1, it was observed that video 1 evokes happiness, neutral to positive valence and high focal attention. Video 2 evoked sadness and disgust, in which the valence was negative, the arousal was medium and there was a low focus of attention. Video 3 evoked mostly sadness, fear and disgust, in which valence was negative, arousal was medium and the focus of attention was

low. Video 4 showed a high level of happiness, positive valence, a medium arousal and high attention. In video 5, the main emotion was sadness, there was negative valence, the level of arousal was high and the focus of attention was high. Finally, the video 6 showed in the first part of video sadness and the last part happiness, with negative and positive valence respectively, a high arousal and a medium focus of attention.

Table 7.8 Recognition of emotions evoked for each video by volunteer 1.1

| Case 1 | Video 1 | Video 2 | Video 3 | Video 4 | Video 5 | Video 6 |
|-----------------|-------------|----------|----------|----------|----------|---------|
| Sadness | - | Yes | Yes | - | Yes | Yes |
| Happiness | Yes | - | - | Yes | - | Yes |
| Fear | - | - | Yes | - | - | - |
| Disgust | - | Yes | Yes | - | - | - |
| Valence | Neutral/pos | Negative | Negative | Positive | Negative | Neg/pos |
| Arousal | Low | Medium | Medium | Medium | High | High |
| Focal Attention | High | Low | Low | High | High | Medium |

Case 2

Table 7.9 indicates the emotions recognized by the multisensorial system. For volunteer 2, it was observed that video 1 evokes happiness, neutral valence, low arousal and high focus of attention. Video 2 evoked happiness and disgust, the valence was neutral, there was low arousal and medium focus of attention. Video 3 evoked happiness and fear, while valence was neutral, there was low arousal and low focus of attention. Video 4 showed a high level of happiness, positive valence, a low arousal and high focus of attention. In video 5, the main emotion was sadness, negative valence, and the level of arousal was medium and the focus of attention was high. Finally, the video 6 showed happiness, with negative and positive valence, medium arousal and high focus of attention.

32 Table 7.9 Recognition of emotions evoked for each video by volunteer 2.

| Case 2 | Video 1 | Video 2 | Video 3 | Video 4 | Video 5 | Video 6 |
|-----------------|---------|---------|---------|----------|----------|---------|
| Sadness | - | - | - | - | Yes | - |
| Happiness | Yes | Yes | Yes | Yes | - | Yes |
| Fear | - | - | Yes | - | - | - |
| Disgust | - | Yes | - | - | - | - |
| Valence | Neutral | Neutral | Neutral | Positive | Negative | Neg/pos |
| Arousal | Low | Low | Low | Low | Medium | Medium |
| Focal Attention | High | Medium | Low | High | High | High |

Case 3

Table 7.10 indicates the emotions recognized by the multisensorial system. For volunteer 3, it was observed that video 1 evokes happiness, positive valence, low arousal and high focus of attention. Video 2 evoked disgust, the valence was neutral, with low arousal and high focus of attention. Video 3 evoked fear, while valence was neutral, with low arousal and medium focus of attention. Video 4 showed happiness, positive valence, medium arousal and high focus of attention. In video 5, the main evoked emotion were sadness and disgust, with negative valence, high level of arousal and medium focus of attention. Finally, video 6 showed sadness, with negative valence, high arousal and low focus of attention.

7.10 Recognition of emotions evoked for each video by volunteer 3.

| Case 3 | Video 1 | Video 2 | Video 3 | Video 4 | Video 5 | Video 6 |
|-----------------|----------|---------|---------|----------|----------|----------|
| Sadness | - | - | - | - | Yes | Yes |
| Happiness | Yes | - | - | Yes | - | - |
| Fear | - | - | Yes | - | - | - |
| Disgust | - | Yes | - | - | Yes | - |
| Valence | Positive | Neutral | Neutral | Positive | Negative | Negative |
| Arousal | Low | Low | Low | Medium | High | High |
| Focal Attention | High | High | Medium | High | Medium | Low |

The three case studies show that the evoked emotions were those proposed in Table 7.7. The facial expressions detected correspond to those expected, with valence levels also related to the expected ones. The levels of arousal were not so high, which may be due to the videos were edited in order to have low negative impact. Finally, the focus of attention detected also corresponded to those expected.

7.4. Discussion

In this research, a multisensorial system for emotions recognition was developed. The system is based on the integration of three sensors (Kinect, eye tracker and thermal camera). The proposed strategy for integration of these sensors was designed to improve the emotion recognition system, which is based on the detection of focus of attention, expression recognition and thermal variation. The technique used for detection of focus of attention was IR-pupil corneal reflection (IR-PCR) introduced in Duchowski (2003) and Bengoechea et al. (2012), which provides highly accurate gaze point measurements, of up to 0.5° of visual angle. For expression recognition, the implementation was based on Facial Action Coding System (FACS), that describes all possible perceivable facial muscle movements in terms of

predefined action units (AUs) proposed by Ekman and Friesen (1978) and implemented in different researches (Mase, 1991; Essa, 1997; and Bartlett et al., 1999). For detection of thermal variation, two techniques were implemented: Facial Thermal - Region of Interest (FT-RoI) introduced in Veltman and Vos (2005) and Khan, Ward, and Ingleby (2009), and Facial Thermal Feature Points (FTFP), which have been tried by Khan et al. (2005), Pavlidis (2004) and Sugimoto (2000).

Different procedures for multisensorial integration have been proposed in the literature (Pantic 2003), (Al Osman, 2016), (Lingenfelser, 2011) and (Wagner, 2011), nevertheless, the multisensorial approach presents challenges related to the integration of individual signals from the different sensors, dimensionality of the feature space, and incompatibility of collected signals in terms of time resolution and format.

In order to integrate Kinect, eye tracker and thermal camera, three integration levels were implemented. In the first stage of the system, eye tracker and Kinect were integrated using a decision-level technique. Feature-level technique was used to integrate thermal camera and Kinect in the second stage. During the third stage, a hybrid-level technique was used to integrate thermal camera, eye tracker and Kinect.

To test and evaluate the multisensorial system, three experiments were proposed the first experimental procedure was designed to evaluate social visual attention. The second procedure was proposed to evaluate the recognition of facial expressions and emotional variation, and the third procedure was designed to evaluate emotions by integrating the three sensors.

Experiment 1 allowed evaluating focal attention and valence comprehension. The results obtained showed that images that correspond to positive valence have the highest percentage of average time of viewing. On the other hand, images that correspond to negative valence have the lowest percentage of average time of viewing, and images with positive valence have the highest number of observers. The images with neutral and negative valence elicited low attention in 8 and 7 participants, respectively, whereas images with positive valence elicited low attention in 1 participant. All these results about focus of attention and comprehension of valence are considered very important for medical and psychological therapies as well as evaluation tool for therapists.

Experiment 2 allowed evaluating facial expressions recognition and emotional variation. The results obtained show that the highest focus of attention is relative to the central regions of the pictures exhibited in the screen, featured by the regions of the eyes, nose and cheeks. On the other hand, the average time required by the participants to recognize the six facial expressions was lower for happiness and higher for disgust. It was also noted that for the sixteen volunteers, the number of mistakes for emotion recognition was higher for disgust while happiness had 0 mistake. Finally, the expression recognition was 70%, and the average focus of attention was 78%. No changes were detected in the temperature that would show variation in valence or arousal. We believed that this is due to the small changes in facial expressions, as the muscle movements are very smooth, and the low sensibility of the thermal

sensor does not allow measuring such small temperature variation, as explained in Section 5.4.

Experiment 3 allowed evaluating valence, arousal and emotion recognition. The three case studies show that the evoked emotions were those proposed in Table 7.7. This experiment only shows a trend, but it is not conclusive because only three selected cases were evaluated from the database, out of 105 children who performed the experiment.

The major difficulties of the experiments were that the data acquisition for the three sensors is not synchronized and a manual synchronization process is required. The characteristics of the thermal camera (vision mode, image resolution and sensitivity) are not suitable for the proposed procedures. On the other hand, the eyes are small and very limited regions to be analyzed. Despite the difficulties presented, the system has potential to be used in applications of emotion recognition, although more investigations are necessary.

Table 7.11 shows the evaluation of functional and technical sensors features presented in Section 1.1.2. According to the results, the system is able to meet all the technical requirements, in case of suitable operation of thermal camera.

Table 7.11 Validation of functional and technical sensors features

| Sensors number | Functional requirements | | | | | | Technical requirements | | | |
|-----------------------|--------------------------------|--------------------------------|-----------------------------------|---------------------------------|------------------------------|-------------------------------|-------------------------------|-------------------|---------------------------|---------------------------------------|
| | 1 Focus of attention | 2 Valence comprehension | 3 Expression comprehension | 4 Expression recognition | 5 Valence recognition | 6 Emotional evaluation | 7 Contactless | 8 Portable | 9 Robust operation | 10 easy to set up, calibration |
| Eye tracker | X | X | X | | | | X | X | X | X |
| Kinect | | | | X | | | X | X | X | X |
| Thermal cam | | | | | X | | X | X | X | X |
| Multisensorial | X | X | X | X | X | X | X | X | X | X |

Finally, it was observed that, once the thermal camera limitation is overcome, the multisensorial system can be used in the evaluation of emotions, and integrated to a robot or computer.

CHAPTER 8

8. CONCLUSIONS AND FUTURE WORKS

8.1. Conclusions:

In this M.Sc. Thesis, the development of a multisensorial system, composed of three sensors, for emotion recognition was introduced. The advantage of such multisensorial system was that the three sensors allowed exploring different emotional aspects, as the eye tracker, using the IR-PCR technique, helped conducting studies about visual social attention; the Kinect, in conjunction with the FACS-AU system technique, allowed developing a tool for facial expression recognition; and the thermal camera, using the FT-RoI technique, was employed for detecting facial thermal variation. When performing the multisensorial integration of the system, it was possible to obtain a more complete and varied analysis of the emotional aspects, allowing evaluate focal attention, valence comprehension, valence expressions, facial expression, valence recognition and arousal recognition.

In Chapter 1, a general review of the different techniques used for automatic recognition of emotions was presented. Various modalities of emotional channels were used for the automated recognition, and each one provides different measurable information to estimate human emotion. In this context, different technologies have been developed to detect human emotional information, and each technology presents advantages and disadvantages, depending on the application. Color camera-based systems continue being the gold standard technology to estimate facial emotions. Eye tracking technologies have emerged as an important tool in recognition of visual social attention and are widely used for research and commercial purposes, while technologies based on thermal device have begun to be studied in the last years. After the bibliographic review of Chapter 1, three devices were proposed to be used in this work: eye tracker, Kinect and thermal camera. These devices have important advantages, since they are contactless (non-invasive), portable, besides having a robust operation and being easy to set up.

The methodological aspects on which the research was based were presented in Chapter 2. The construction of an experimental platform allowed the integration of the devices in a box, facilitating the transportation and adaptation of the platform to different experimental environments. These experimental environments were previously adapted in conditions of light, temperature, humidity and noise required for the tests, and filters were used to attenuate external factors that could affect the results of the research.

The implementation of an eye tracking interface was presented and validated in Chapter 3. An eye tracker was used to identify eye gaze, in order to recognize the visual focal attention of a person. The main problem was that there is no interface to connect the eye tracker to an application in Matlab. To solve this problem, a server in Python was used and an interface in Matlab was developed. This interface was very important for this study, since it allowed automating the experiments of visual attention required to evaluate emotions and, additionally, the use of eye tracking technique in other assistive applications of our lab, such as controlling intelligent environments, wheelchairs, intelligent walkers, etc.

In Chapter 4, a system for facial expression recognition using the Kinect was presented. Detecting facial features for expressions recognition is a difficult task, and, in order to fulfill this objective, a method based on the FACS-AU facial muscle system was implemented, using the Brekel software to obtain the AU face features. Then, KNN and LDA algorithms were implemented to recognize the six basic facial expressions. The results obtained reached about 70% of success rate. This low success rate is due to the system is based on Brekel, which only allows the detection of 20 AUs, while the FACS system has more than 44 AUs. However, the results show the possibility to implement algorithms to detect more AUs , and, consequently, improve the accuracy.

A facial thermal variation detection method was presented in Chapter 5. Detection of emotions using the modality of thermal physiological variation is one of the most controversial in the literature, since representative model for estimating the relation between fluctuations in facial temperature and facial emotional activity is not yet available. In this work, two approaches were studied: in the first one, the thermal facial variation was evaluated based on the analysis of variation in the facial expression. Here, due to the features of the thermal sensor, which does not have the required sensitivity, no thermal variation was detected. Nevertheless, in the second approach, based on thermal variation related to changes in arousal and emotional valence, it was possible to measure thermal facial variations that correspond to changes in emotions. These variations were detected in the nose, cheeks, and forehead. The results obtained are not conclusive and the use of a thermal camera with better performance is required.

A multisensorial integration strategy was presented in Chapter 6. The multisensorial approach presents challenges related to the fusion of individual signals from the different sensors, dimensionality of the feature space, and incompatibility of collected signals in terms of time resolution and format. The strategies presented in this work allowed to integrate such heterogeneous devices like eye tracker, Kinect and thermal camera into a all-in-one system. The main difficulties in the integration were the range of operating of each equipment and the difficulty of synchronizing the data that were captured by different computers. These problems were solved, firstly by changing the setup for the experimental tests, and secondly

with a manual synchronization of the videos. The integration of eye tracker and Kinect allowed to perform joint studies of focus of attention in recognition of facial expressions and valence, while the Kinect-thermal camera integration allowed proposing a novel technique using the AU in the thermal images, which has improved the detection and segmentation of FT-RoI in the thermal image.

In Chapter 7, the multisensorial system was validated. The multisensorial system was tested in sixteen adults and three children volunteers. An experimental protocol for evoking emotions was proposed to be used with the developed system, which was able to detect eye gaze, recognize facial expression and estimate the valence and arousal for emotion recognition, fulfilling the main objective of this M.Sc. Thesis.

Finally, with the system here developed, emotions of people can be analyzed by facial features using contactless sensors in semi-structured environments, such as clinics, laboratories, or classrooms. This system also presents the potential to become an embedded tool in robots to endow these machines with an emotional intelligence for a more natural interaction with humans.

8.2. Contributions

The main contribution of this research was the development of a multisensorial system in order to automate the emotion recognition. The system allows detecting visual social attention, recognizing facial expression, estimating the valence emotion, and integrating results for further evaluation. The calibration procedure is fast and performed at the beginning of each experiment. Also, the integrated system provides an easy-to-use tool, which is versatile, robust, contactless and portable, and, additionally, can be used in social emotion therapy and assistive robotic applications. Other contributions involve the development of an eye tracking interface for assistive applications using an eye tracker as a tool for social visual attention applications and control of devices through eye gaze. In addition, this research presents a novel technique for FACS-AU and FT-RoI integration in order to improve the detection and segmentation of FT-RoI in thermal image. Additionally, a database of more than 100 children with facial information from color camera, thermal camera and eye tracking in a semi-controlled environment was collected. This is an important contribution, since, in the literature, it is difficult to find a multisensorial database of children's emotional information.

8.3. Publications

During this research, the following publications were realized:

- **RIVERA, H.**; GOULART, C.; CALDEIRA, E.; BASTOS, T. Using Eye-Tracking for the Study about Valence and Emotional Facial Expressions. In: Anais do XXV Congresso Brasileiro de Engenharia Biomédica CBEB 2016.
- **RIVERA, H.**; COTRINA, A.; VALADAO, C.; BENEVIDES, A.; BASTOS, T. Motor Intention Detection for Robotic Walker Users Using Artificial Neural Networks and Eye-Tracking. In: Anais do XXV Congresso Brasileiro de Engenharia Biomédica CBEB 2016.
- **RIVERA, H.**; BISSOLI, A.; GOULART, C.; CALDEIRA, E.; BASTOS, T. Development of Matlab Toolbox for Eye Tracking Systems. In: Anais do XXI Congresso Brasileiro de Automática CBA 2016.
- GOULART, C.; **RIVERA, H.**; FAVARATO, A.; BINOTTI, V.; BALDO, G.; VALADAO, C.; CALDEIRA, E.; BASTOS, T. Towards an Improved Human-Affective Robot Interaction. In: Anais do XXV Congresso Brasileiro de Engenharia Biomédica CBEB 2016.
- COTRINA, A.; VALADAO, C.; **RIVERA, H.**; BENEVIDES, A.; BASTOS, T. Towards Motor Intention Detection of Robotic Walker Users Based on Brain-Computer Interfaces. In: Anais do XXV Congresso Brasileiro de Engenharia Biomédica CBEB 2016.
- VALADAO, C.; GOULART, C.; **RIVERA, H.**; CALDEIRA, E.; BASTOS, T.; FRIZERA NETO, A.; CARELLI, R. Analysis of the use of a robot to improve social skills in children with autism spectrum disorder. In: Research on Biomedical Engineering RBE. 2016.
- GOULART, C.; **RIVERA, H.**; VALADAO, C.; CALDEIRA, E.; BASTOS, T. Recognizing Emotions and Focus of Attention in Individuals with ASD Based on Facial Images. In: Anais do VI Congresso Brasileiro de Biotecnologia 2015.
- COTRINA, A.; Glasgio, G.; **RIVERA, H.**; Ferreira, A.; BASTOS, T. Evaluation of Eye Movements Using Tracking Devices in Context of a Novel Ssvep-Bci Setup. In: Anais do XII Simpósio Brasileiro de Automação Inteligente SBAI 2015.

8.4. Future works

The following tasks are indicated as possible future works:

- Synchronize the database information and testing algorithms in order to improve the emotion detection and classification.
- Try a thermal camera with higher resolution and better sensitivity to capture thermographic images suitable for detecting emotions
- Implement methods to detect more AUs in order to improve the facial expression detection and classification.
- Develop a strategy of synchronization of the sensors that allows to work on-line and detect emotions in real time.
- Integrate the emotions recognition system to a multimedia system (animated face, sound and video) to create affective computing applications.
- Test the multisensory system in experimental therapy with children with autism.

REFERENCES

- AL OSMAN, H.; DONG, H.; EL SADDIK, A. Ubiquitous biofeedback serious game for stress management. *IEEE Access*, vol. 4, pp. 1274–1286, 2016.
- ALEKSIC, P. S.; KATSAGGELOS, A. K. Automatic facial expression recognition using facial animation parameters and multistream hmms. *TIFS*, vol. 1, no. 1, pp. 3–11, 2006.
- ALEXANDRE, L. A.; CAMPILHO, A. C.; KAMEL, M. On combining classifiers using sum and product rules. *Pattern Recognition Letters*, vol. 22, pp. 1283–1289, 2001.
- ALYUZ, N.; GOKBERK, B.; AKARUN, L. Adaptive registration for occlusion robust 3D face recognition. *ECCV*, 2012.
- AMBADY, N.; ROSENTHAL, R. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, vol. 111, p. 256, 1992.
- BÄNZIGER, T.; MORTILLARO, M.; SCHERER, K. R. Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, vol. 12, p. 1161, 2012.
- BARTLETT, M.; HAGER, J.; EKMAN P.; SEJNOWSKI, T. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–264, 1999.
- BARTLETT, M.; LITTLEWORT, G.; FRANK, M.; LAINSCSEK, C.; FASEL, I.; MOVELLAN, J. Recognizing facial expression: Machine learning and application to spontaneous behavior. *Proc. of IEEE Conf. on Computer Vision and Pat. Recog. (CVPR)*, pp. 568–573, 2005.
- BENGOCHEA, J.; VILLANUEVA, A., CABEZA, R. Hybrid eye detection algorithm for outdoor environments. *Proceedings of the 2012 ACM conference on ubiquitous computing, UbiComp'12*. ACM, New York, pp 685–688, 2012
- BERNHARDT, D.; ROBINSON, P. Detecting affect from non-stylised body motions. *International conference on affective computing and intelligent interaction*, 2007.
- BLOM, P. M.; BAKKES, S.; TAN, C.T.; WHITESON, S.; ROIJERS, D.; VALENTI, R.; GEVERS, T. Towards personalised gaming via facial expression recognition. *AIIDE*, 2014.
- BORGHETTI, D.; BRUNI, A.; FABBRINI, M. A low-cost interface for control of computer functions by means of eye movements. *Comput Biol Med* 37(12):1765–1770, 2007.

- BRIESE, E.; CABANAC, M. Stress hyperthermia: Physiological arguments that it is a fever. *Physiological Behavior*, 49, 1153–1157, 1991.
- BURKE, J.; MCNEILL, M.; CHARLES, D.; MORROW, P.; CROSBIE, J.; MCDONOUGH, S. Optimising engagement for stroke rehabilitation using serious games. *The Visual Computer* December 2009.
- C#. User Manual, Available at: <https://www.microsoft.com/net>, 2013.
- CAMURRI, A., LAGERLÖF, I., VOLPE, G. Recognizing Emotion from Movement: Comparison of Spectator Recognition and Automated Techniques, *International Journal of Human-Computer Studies*, 59(1-2), pp. 213-225, Elsevier Science, 2003.
- CHARLES DARWIN. *The Expression of the Emotions in Man and Animals*, England, 1904.
- COLOMBO, A.; CUSANO, C.; SCHETTINI, R. 3D face detection using curvature analysis. *PR*, vol. 39, no. 3, pp. 444–455, 2006.
- DALGLEISH, T.; DUNN, B.; DMOBBS D. Affective neuroscience: Past, present, and future. *Emotion Review*, vol. 1, pp. 355–368, 2009.
- DARWIN C. *The expression of the emotions in man and animals*, London, UK: John Murray, 1965.
- DARWIN, C. *The Expression of the Emotions in Man and Animals*, 3rd edit. Introduction, afterwords, and commentaries by Paul Ekman. Harper Collins. London (US edit.: Oxford University Press. New York), 1872.
- DELAC, K.; GRGIC, M.; GRGIC, S. Independent comparative study of PCA, ICA, and LDA on the FERET data set. *Int. J. Imag. Syst. Technol.*, vol. 15, no. 5, pp. 252–260, 2005.. Available at: <http://www.image.ntua.gr/ermis/>.
- DEVAULT, D.; ARTSTEIN, R.; BENN, G.; DEY, T.; FAST, E.; GAINER, A.; MORENCY, L. A virtual human interviewer for healthcare decision support.” *AAMAS*, 2014.
- DHALL, A.; GOECKE, R.; LUCEY, S.; GEDEON, T. Collecting large, richly annotated facial- expression databases from movies. *IEEE Multimedia*, vol. 19, pp. 34–31, 2012.
- DONATO, G.; BARTLETT, M.; HAGER, J.; EKMAN, P.; SEJNOWSKI, T. Classifying facial actions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.
- DOUGLAS-COWIE, E.; COWIE, R.; SCHROEDER, M. The description of naturally occurring emotional speech. *Proc. 15th Int. Conf. Phonetic Sciences*, Barcelona, Spain, 2003.
- DOUGLAS-COWIE, E.; COWIE, R.; SNEDDON, I.; COX, C.; LOWRY, O.; MCRORIE, M. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. *Inte* 2011.

- DU, S.; TAO, Y.; MARTINEZ, A. Compound facial expressions of emotion. *Proc Natl Acad Sci.*; 1 (11): E1454–E1462, 2014.
- DUCHOWSKI, A. T.; VERTEGAAL, R. Eye-based interaction in graphical systems: theory and practice. Course 05, SIGGRAPH 2000. ACM, New York, 2000.
- EKMAN P. *Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional Life.* Times Books, USA, First edition, 2003.
- EKMAN, P. *Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional Life.* Times Books, USA, First edition, 2003.
- EKMAN, P. Facial expression and emotion. *American Psychologist*, 48:384–392, 1993.
- EKMAN, P. FRIESEN, W. *The Facial Action Coding System: A Technique For The Measurement of Facial Movement.* Consulting Psychologists Press, Inc., San Francisco, CA, 1978.
- EKMAN, P.; Friesen, W. *Pictures of facial affect.* Palo Alto, CA: Consulting Psychologist, 1976.
- ESSA, I.; PENTLAND, A. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 19(7):757–763, 1997.
- EYBEN, F.; WÖLLMER, M.; SCHULLER, B. OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit 3rd international conference on affective computing and intelligent interaction and workshops, 2009.
- EYE TRIBE SERVER FOR MATLAB. User Manual, Available at: <https://github.com/esdalmajer/EyeTribe-Toolbox-for-Matlab>, 2013
- FASEL, B.; LUETTIN, J. Automatic facial expression analysis: A survey,” *Pattern Recognition*, vol. 36, pp. 259–275, 2003.
- FRIESEN, W. V.; EKMAN, P. *Emfacs-7: Emotional facial action coding system.* U. California, vol. 2, p. 36, 1983.
- GAJSEK, R.; STRUC, V.; MIHELIC, F.; PODLESEK, A.; KOMIDAR, L.; SOCAN, G.; BAJEC, B. Multi-modal emotional database: AVID. *Informatica* 33, pp. 101–106, 2009.
- GELDER DE B. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, pp. 3475–3484, 2009.
- GENNO, H.; ISHIKAWA, K.; KANBARA, O.; KIKUMOTO, M.; FUJIWARA, Y.; SUZUKI, R.;. Using facial skin temperature to objectively evaluate sensations. *International Journal of Industrial Ergonomics*, 19(2), 161–171, 1997.
- GOLDBERG, J. H.; WICHANSKY, A. M. Eye tracking in usability evaluation: a practitioner’s guide. In: Hyönä J, Radach R, Deubel H (eds) *The mind’s eye: cognitive*

- and applied aspects of eye movement research. North-Holland, Amsterdam, pp 493–516, 2003.
- GONG, L.; WANG, T.; WANG, C., LIU, F., ZHANG, F., YU X., Recognizing affect from non-stylized body motion using shape of Gaussian descriptors. Proceedings of the 2010 ACM symposium on applied computing, 2010
- GOULART, C.; CASTILLO-GARCIA, J.; VALADÃO, C.; CALDEIRA, E.; BASTOS-FILHO, T. Study of EEG Signals to Evaluate Emotional and Mental States of children with ASD in the Interaction with Mobile Robot. International Workshop on Assistive Technology (IWAT), 2015.
- GUNES, H.; PICCARDI, M. Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications, 2006.
- GUPTA, R.; BANVILLE, H.; FALK, T. PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience. IEEE workshop on applications of signal processing to audio and acoustics (WASPAA), 2015.
- GUPTA, R.; FALK, T. H. Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization. Neurocomputing, vol. 174, pp. 875–884, 2016.
- GUPTA, R.; KHOMAMI, M.; CÁRDENES, J. ; MORREALE, F.; FALK, T.; AND SEBS, N. A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. Proceedings of the 2016 ACM on international conference on multimedia retrieval, 2016.
- HANSEN, D. W.; MAJARANTA, P. Basics of camera-based gaze tracking. In: Majaranta P et al (eds) Gaze interaction and applications of eye tracking: advances in assistive technologies. Medical Information Science Reference, Hershey, pp 21–26, 2012.
- HANSEN, D. W.; Pece, A. Eye tracking in the wild. Comput Vis Image Underst 98(1):155–181, 2005.
- HEALEY, J. A.; PICARD, R. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems, vol. 6, pp. 156–166, 2005.
- HOMMA, I.; MASAOKA, Y. Breathing rhythms and emotions. Experimental Physiology, vol. 93, pp. 1011–1021, 2008.
- HOOKE, K. Knowing, Communicating, and Experiencing through Body and Emotion. IEEE Transactions on Learning Technologies, 2008.
- HORI, J.; SAKANO, K.; MIYAKAWA, M.; SAITOH, Y. Eye movement communication control system based on EOG and voluntary eye blink. Proceedings of the 9th international conference on computers helping people with special needs, ICCHP, vol 4061, pp 950–953, 2006.

- HUGHES, G. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, vol. 14, pp. 55–63, 1968
- IRANI, R.; NASROLLAHI, K.; SIMON, M. O.; CORNEANU, C. A.; ESCALERA, S.; BAHNSEN, C.; LUNDTOFT, D. H.; MOESLUND, T. B.; PEDERSEN, T.; KLITGAARD, M. Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition. *CVPR Workshops*, 2015.
- IRIS. Available at: <http://www.cse.ohiostate.edu/OTCBVSBENCH/Data/02/download.html>. 2010.
- ISHIGURO, H.; ONO, T.; IMAI, M.; MAEDA, T.; KANDA, T.; NAKATSU, R. Robovie: an interactive humanoid robot. *Industrial robot: An international journal*, vol. 28, no. 6, pp. 498–504, 2001.
- IZARD, C. E. A system for identifying affect expressions by holistic judgments. *Instructional Resources Center, University of Delaware*, 1983.
- IZARD, C. E. Maximally discriminative facial movement coding system (MAX). *Instructional Resources Center, University of Delaware*, 1983.
- IZARD, C.; DOUGHERTY, L.; HEMBREE, E. A system for identifying affect expressions by holistic judgments. *Unpublished Manuscript, University of Delaware*, 1983.
- JERRITTA, S.; MURUGAPPAN, M.; NAGARAJAN, R.; WAN, K. Physiological signals based human emotion recognition. a review in 2011 *IEEE 7th international colloquium on signal processing and its applications (CSPA)*, 2011.
- JIawei, HAN.; MICHELINE, K.; JIAN, P. *Data Mining: Concepts and Techniques*. Elsevier, USA, Third edition 2012.
- JOVANOVIĆ, E.; LORDS, A.; RASKOVIĆ, D.; COX, P.; ADHAMI, R.; ANDRASIĆ, F. Stress monitoring using a distributed wireless intelligent sensor system. *IEEE Engineering in Medicine and Biology Magazine*, vol. 22, pp. 49–55, 2003.
- KAPOOR, A.; BURLESON, W.; PICARD, R. W. Automatic prediction of frustration, *IJHCS*, vol. 65, no. 8, pp. 724–736, 2007.
- KHAN, M. M.; WARD, R. D.; INGLEBY, M. Automated classification and recognition of facial expressions. In *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, Singapore, (Dec), 202–206. 2004
- KHAN, M. M.; WARD, R. D.; INGLEBY, M. Infrared thermal sensing of positive and negative affective states. In *Paper presented at the conference on robotics, automation and mechatronics*, Bangkok, 2006.
- KHAN, M. M.; WARD, R. D.; INGLEBY, M. The distinguishing facial expressions by thermal imaging using facial thermal feature points. In *Proceedings of the 19th British HCI Group Annual Conference (HCI'05)*, Edinburgh, (Sept), L. Mackinnon, O.

- Bertelsen and N. Bryan-Kinns Eds. The British Computer Society, London, UK. 10–14. 2005.
- KHAN, M. M.; WARD, R.; INGLEBY, M. Infrared Thermal Sensing of Positive and Negative Affective States, Robotics, Automation and Mechatronics. IEEE Conference on, pp.1-6, Dec. 2006.
- KHAN, M.M.;WARD, R.; INGLEBY, M. Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature, Trans. Appl. Percept., vol.6, no. 1, pp. 1–22, 2009.
- KIM, J.; ANDRÉ, E.; REHM, M.; VOGT, T.; WAGNER, J. Integrating information from speech and physiological signals to achieve emotional sensitivity. Proc. INTERSPEECH, Lisboa, Portugal, 2005.
- KINECT 2.0. User Manual, Available at: <https://developer.microsoft.com/en-us/windows/kinect/develop>, 2015.
- KINECT SDK. User Manual, 2013 Available at: “Programming Guide: Face Tracking”, <http://msdn.microsoft.com/en-us/library/jj130970.aspx>, Microsoft MSDN, 2013..
- KLEINSMITH, A.; BIANCHI-BERTHOUBE, N. Recognizing affective dimensions from body posture. International conference on affective computing and intelligent interaction, 2007.
- KOESLTRA, S.; MÜHL, C.; SOLEYMAN, M.; LEE, J. S.; YAZDANI, A.; EBRAHIMI, T.; PUN, T.; NIJHOLT, A.; PATRAS, I. DEAP: A database for emotion analysis using physiological signals. Transactions on affective computing”, vol 3, no 1, 2012.
- KURAOKA, K.; NAKAMURA, K. The use of nasal skin temperature measurements in studying emotion in macaque monkeys. Physiology Behaviour, 1(102), 347–355, 2011.
- LANG, P. J. The emotion probe: Studies of motivation and attention. American Psychologist, 50, 371– 385, 1995.
- LANG, P. J.; BRADLEY, M. M.; CUTHBERT, B. N. International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual (Technical Report No. A-6). Gainesville, FL: University of Florida, Center for Research in Psychophysiology, 2005
- LANG, P. J.; BRADLEY, M. M.; CUTHBERT, B. N. International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-. University of Florida, Gainesville, FL. 2008.
- LIEN, J.; KANADE, T.; COHN, J.; LI, C. Detection, tracking, and classification of action units in facial expression. Journal of Robotics and Autonomous System, 31:131–146, 2000.
- LIN, J.-C.; WU, C.-H.; WEI, W. L. Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. IEEE Transactions on Multimedia, vol. 14, pp. 142–156, 2012.

- LINGENFELSER, F.; WAGNER, J.; ANDRÉ, E. A systematic discussion of fusion techniques for multi-modal affect recognition tasks. Proceedings of the 13th international conference on multimodal interfaces, 2011.
- MAAT, L.; PANTIC, M. Gaze-x: adaptive, affective, multimodal interface for single-user office scenarios. Artificial Intelligence for Human Computing. Springer, 2007.
- MAO, Z.; SIEBERT, J.; COCKSHOTT, W.; Ayoub, A. Constructing dense correspondences to analyze 3D facial change. ICPR, 2004.
- MARTIN, O.; KOTSIA, I.; MACQ, B.; PITAS, I. The eNTERFACE'05 audio-visual emotion database. 22nd international conference on data engineering workshops (ICDEW'06), 2006.
- MASE, K.; Recognition of facial expression from optical flow. IEICE Transactions, E. 74(10):3474–3483, 1991.
- MATLAB. User Manual, 2013 Available at: <https://www.mathworks.com/products/matlab>
- MCDUFF, D.; GONTAREK, S.; PICARD, R. Improvements in remote cardiopulmonary measurement using a five band digital camera. IEEE Transactions on Biomedical Engineering, vol. 61, pp. 2593–2601, 2014.
- MCNEILL, D. Hand and mind: What gestures reveal about thought, Chicago, IL: University of Chicago Press, 1992.
- MICHAUD, F.; CLAVET, A. Robotoy contest — designing mobile robotic toys for autistic children. Proceedings of The American Society for Engineering Education (ASEE'01), Albuquerque, 2001.
- MURUGAPPAN, M. RAMACHANDRAN, N. SAZALI, Y. Classification of human emotion from EEG using discrete wavelet transform". J. Biomedical Science and Engineering, vol. 3, 390-396, 2010.
- NAGUMO, K.; ZENJU, H.; NOZAWA, A.; IDE, H.; TANAKA, H. Evaluation of temporary arousal level using thermogram images. In Paper presented at the 19th remote sensing forum, 3 March, Tokyo, Japan, 2002.
- NAIR, P.; CAVALLARO, A. 3-d face detection, landmark localization, and registration using a point distribution model. T.Multimedia, vol. 11, no. 4, pp. 611–623, 2009.
- NAKANISHI, R.; IMAI-MATSUMURA, K. Facial skin temperature decreases in infants with joyful expression. Infant Behavior & Development, 31, 137–144. <http://dx.doi.org/10.1016/j.infbeh.2007.09.001>, 2008.
- NAKASONE, A.; PRENDINGER, H.; ISHIZUKA, M. Emotion recognition from electromyography and skin conductance. Proc. of the 5th international workshop on biosignal interpretation, 2005.
- NIST Equinox Available at: <http://www.equinoxsensors.com/products/HID.html>.

- O'TOOLE, A. J.; HARMS, J.; SNOW, S. L.; HURST, D. R.; PAPPAS, M. R.; AYYAD, J. H.; ABDI, H. A video database of moving faces and people. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 812–816, 2005.
- PANNING, A.; SIEGERT, I.; AL-HAMADI, A.; WENDEMUTH, A.; RÖSNER, D.; FROMMER, J. Multimodal affect recognition in spontaneous hci environment. *IEEE international conference on signal processing, communication and computing (ICSPCC)*, 2012.
- PANTIC, M., ROTHKRANTZ, L.J.M. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
- PANTIC, M., ROTHKRANTZ, L.J.M. Towards an Affect-sensitive Multimodal Human-Computer Interaction, *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-13902, 2003.
- PANTIC, M.; ROTHKRANTZ, L. J. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, vol. 91, pp. 1370–1390, 2003.
- PANTIC, M.; STEWART-BARTLETT, M. Machine Analysis of Facial Expressions, in *Face Recognition*, K. D. a. M. Grgic, Ed. Vienna, Austria: I-Tech Education and Publishing, 2007.
- PAVLIDIS, I.; EBERHARDT, N. L.; LEVINE, J. A. Human behaviour: Seeing through the face of deception. *Nature*, 4, 35. <http://dx.doi.org/10.1038/415035a>, 2002.
- POH, M.; MCDUFF, D.; PICARD, R. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 7–11, 2011.
- PROCESSING. User Manual, Available at: <https://processing.org>, 2014.
- PYTHON. User Manual, 2013 Available at: <https://docs.python.org/3/tutorial/index.html>
- RIMM-KAUFMAN, S. E.; KAGAN J. The psychological significance of changes in skin temperature. *Motivation and Emotion*, vol. 20, pp. 63–78, 1996.
- ROBINS, B.; DAUTENHAHN, K.; TE BOEKHORST, R.; BILLARD, A. Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills. *Univ. Access Inf. Soc.*, vol. 4, p. 105-120, 2005.
- ROISMAN, G. I.; TSAI, J. L.; CHIANG, K. S. The emotional integration of childhood experience: Physiological, facial expressive, and self-reported emotional response during the adult attachment interview. *Development. Psychol.*, vol. 40, no. 5, pp. 776–789, 2004.
- ROWLEY, H.; BALUJA, S.; KANADE, T. Neural network-based face detection. *IEEE Transactions On Pattern Analysis and Machine intelligence*, 20(1):23–38, 1998.

- RYAN, A.; COHN, J.; LUCEY, S.; SARAGIH, J.; LUCEY, P.; TORRE, F.; ROSS, A. Automated facial expression recognition system. ICCST, 2009.
- SALAZAR-LÓPEZ, E. The mental and subjective skin: Emotion, empathy, feelings and thermography. *Consciousness and Cognition*, 34, 149-162. 2015.
- SAVVA, N.; SCARINZI, A.; BIANCHI-BERTHOUSE, N. Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *IEEE Transactions on Computational Intelligence and AI in games*, vol. 4, pp. 199–212, 2012.
- SCASSELATI, B.; ADMONI, H.; MATARIC, M. Robots for use in Autism Research. *Annual Review of Biomedical Engineering*, 14, 275-294, 2012.
- SCHERER, K.; CESCHI, G. Lost luggage emotion: A field study of emotion-antecedent appraisal. *Motivation and Emotion*, vol. 21, pp. 211–235, 1997.
- SCHERER, K.; EKMAN, P.; *Handbook of methods in nonverbal behavior research*, Cambridge, UK, Cambridge University, 1982.
- SCHERER, K. Adding the affective dimension: A new look in speech analysis and synthesis, *Proc. International Conf. on Spoken Language Processing*, pp. 1808–1811, 1996.
- SCHERER, S.; STRATOU, G.; MAHMOUD, M.; BOBERG, J.; GRATCH, J.; RIZZO, A.; MORENCY, L. Automatic behavior descriptors for psychological disorder analysis. *Automatic Face and Gesture Recognition (FG)*, 10th IEEE International Conference and Workshops on. IEEE, 2013.
- SCHULLER, B.; RIGOLL, G. Recognising interest in conversational speech-comparing bag of frames and supra-segmental features. *Proc. INTERSPEECH*, Brighton, UK, pp. 1999–2002, 2009.
- SEBE, N.; LEW, M.; COHEN, I.; SUN, Y.; GEVERS, T.; HUANG, T. Authentic facial expression analysis. *Proc. 6th IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2004.
- SUGIMOTO, Y.; YOSHITOMI, Y.; TOMITA, S. A method of detecting transitions of emotional states using a thermal facial image based on a synthesis of facial expressions. *Robotics Autonom. Syst.* 31, 147–160, 2000.
- TANAKA, H.; IDE, H.; NAGASHIMA, Y. Attempt of feeling estimation by analysis of nasal skin temperature and arousal level. *Transaction of Human Interface Society*, 1, 51–56, 1999.
- THOMEER, M. L.; SMITH, R. A.; LOPATA, C.; VOLKER, M. A.; LIPINSKI, A. M.; RODGERS, J. D.; MCDONALD, C. A.; LEE, G. K. Randomized Controlled Trial of Mind Reading and In Vivo Rehearsal for High-Functioning Children with ASD. *J Autism Dev Disord.* 13p, 2015.
- ULTRAVNC. User Manual, 2016 Available at: <http://www.uvnc.com/>

- VAM. DATA BASE [Online]. Available: <http://emotion-research.net/download/vam>.
- VELTMAN, J. A.; VOS, W. K. Facial temperature as a measure of operator State. In Paper presented at the 11th international conference on human– computer interaction, 22–27 July, Las Vegas-Nevada, USA, 2005.
- VERMUN, K., SENAPATY, M., SANKHLA, A., PATNAIK, P. ROUTRAY A. Gesture-based affective and cognitive states recognition using kinect for effective feedback during e-learning. IEEE fifth international conference on technology for education (T4E), 2013.
- VURAL, E.; CETIN, M.; ERCIL, A.; LITTLEWORT, G.; BARTLETT, M.; MOVELLAN, J. Drowsy driver detection through facial movement analysis. Human–Computer Interaction, 2007.
- WAGNER, J.; ANDRE, E.; LINGENFELSER, F.; KIM, J. Exploring fusion methods for multimodal emotion recognition with missing data. IEEE Transactions on Affective Computing, vol. 2, pp. 206–218, 2011.
- WENINGER, F.; WÖLLMER, M.; SCHULLER, B. Emotion recognition in naturalistic speech and language—a survey. Emotion Recognition: A Pattern Analysis Approach, Hoboken, NJ: John Wiley & Sons, Inc., pp. 237–267, 2015.
- WERRY, I.; DAUTENHAHN, K.; HARWIN, W. Evaluating the response of children with autism to a robot. Proceedings of Rehabilitation Engineering and Assistive Technology Society of North America, 2001.
- YACOOB, Y.; DAVIS L. Recognizing human facial expression from long image sequences using optical flow. IEEE Trans. on Pattern Analysis and Machine Intell., 18(6):636–642, 1996.
- YOSHITOMI, Y.; KIM, S. I.; KAWANO, T.; KITAZOE, T. Effects of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In Proceedings of the IEEE International Workshop on Robotics and Human Interactive Communication, Osaka, Japan, (Sept), 178–183. 2000.
- ZENJU, H.; NOZAWA, A.; TANAKA, H.; IDE, H. Estimation of unpleasant and pleasant states by nasal thermogram. IEEJ Transactions on Electronics, Information and Systems, 124(1), 213–214. 2004.